eISSN 2300-262X
ISSN 0137-5075

POLISH ACADEMY OF SCIENCES
INSTITUTE OF FUNDAMENTAL TECHNOLOGICAL RESEARCH
COMMITTEE ON ACOUSTICS

ARCHIVES of ACOUSTICS

QUARTERLY

Vol. 51, No. 2, 2026

WARSAW

PAN
POLISH ACADEMY OF SCIENCES
INSTITUTE OF FUNDAMENTAL TECHNOLOGICAL RESEARCH
COMMITTEE ON ACOUSTICS

ARCHIVES of ACOUSTICS

QUARTERLY, Vol. 51, No. 2, 2026

Research Papers

Technical Notes

OSA 2025

Polish Academy of Sciences
Institute of Fundamental Technological Research PAS
Committee on Acoustics PAS

Editorial Office
Institute of Fundamental Technological Research PAS, Pawińskiego 5B, 02-106 Warsaw, Poland
akustyka@ippt.pan.pl https://acoustics.ippt.pan.pl

Indexed in BazTech, Science Citation Index-Expanded (Web of Science Core Collection), ICI Journal Master List, Scopus, PBN – Polska Bibliografia Naukowa, Directory of Open Access Journals (DOAJ)
Recognised by The International Institute of Acoustics and Vibration (IIAV)

Edition co-sponsored by the Ministry of Science and Higher Education

PUBLISHED IN POLAND

Typesetting in LaTeX: Jezierska Katarzyna (Institute of Fundamental Technological Research PAS, Poland)

Archives of Acoustics Vol. 51, No. 2, pp. 235–247 (2026), https://doi.org/10.24423/archacoust.2026.4408

Research Paper

Construction Method of Sparse Dictionary for Multi-Order FRFT Domain Feature Fusion

Tongjing Sun*, Lei Chen, Xiaohong Deng

Department of Automation, Hangzhou Dianzi University
Hangzhou, China

*Corresponding Author: stj@hdu.edu.cn

Received January 5, 2026; accepted March 2, 2026;
available online March 10, 2026; version of record April 24, 2026; published issue June 24, 2026.

Reverberation constitutes a primary source of interference for active sonar signals, particularly the intense reverberation originating from reflections of the incident signal. Sharing the same generation mechanism as the target echo, it severely hampers the extraction and analysis of the target signal. To enhance signal processing capabilities under strong reverberation, this paper proposes a sparse dictionary construction method based on multi-order fractional Fourier transform (FRFT) domain feature fusion. This method exploits the distinctive characteristics exhibited by target echoes and strong reverberation signals across different fractional transform domains to discriminate between them. It constructs sparse sub-dictionaries using these distinct fractional orders, trains the weights of each sub-dictionary via an adaptive gradient optimization strategy to achieve sparse representation of the signal, suppresses strong interference in the sparse domain, and reconstructs the target signal through a reconstruction process, thereby achieving the goal of extracting the target signal while suppressing strong interference. Results from processing lake trial data demonstrate that the proposed method can effectively extract target echo signals amidst strong reverberation, with the signal-to-reverberation ratio improvement consistently no less than 2.1 dB and reaching up to 15.6 dB. This method provides an effective approach for the processing and analysis of weak underwater signals.

Keywords: underwater acoustic signal processing, fractional Fourier transform, feature fusion, sparse reconstruction, strong interference suppression.

1. Introduction

With the continuous advancement of marine exploration technology, active sonar systems play an increasingly vital role in underwater target detection, recognition, and localization (Cui *et al.*, 2023). However, the actual marine environment is complex and variable, containing various noise sources such as waves, wind-generated surface noise, and marine life, all of which interfere with the analysis of target echoes in sonar signals (Yang, 2023). Among these, reverberation stands as a primary background interference for active sonar, especially strong reverberation induced by the transmitted signal. Its spectral structure exhibits a certain similarity to the transmitted signal, rendering conventional time-domain filtering, frequency-domain suppression, and matched filtering methods inadequate for reverberation cancellation (Deng *et al.*, 2005). Consequently, under conditions of strong reverberation interference, traditional methods struggle to effectively extract target echoes, which constitutes a key bottleneck limiting performance enhancement in sonar systems.

Sparse representation theory, which describes signals succinctly using a small number of atoms from an overcomplete dictionary, can better reveal, distinguish, and extract the informational features inherent in signals. This offers a new perspective and methodology for extracting underwater target signals (Ge, Zhang, 2023).

In the past, scholars have conducted related research in this area. Lei (2014) leveraged the structural characteristics of multi-layer wavelet decomposition in conjunction with wavelet modulus maxima line search theory to achieve sparse signal representation based on wavelet modulus maxima search. Sun *et al.* (2016), addressing the issue of processing weak underwater echo signals, proposed a method for constructing an overcomplete atomic library based on prior information, building a library tailored to the signal's own characteristics to achieve sparse reconstruction of underwater echo signals. Liu and Zhou (2018) constructed a joint dictionary for time-frequency overlapping communication signals to perform sparse representation and interference suppression. Sparse representation methods are also widely applied in tasks such as underwater acoustic signal denoising, signal direction finding, and target classification (Zhou *et al.*, 2023; Wu *et al.*, 2021; Meng, 2024; Liao *et al.*, 2014; Wang, 2020). These methods typically construct dictionaries using wavelet bases or the incident signal itself. The principle behind interference suppression relies on the correlation between the echo signal and the incident wave, while noise or other interferences are uncorrelated with it. This allows environmental noise or random scatterers to be filtered out. However, for strong reverberation that shares the same generation mechanism as the target echo signal, it becomes difficult to achieve target signal extraction. To suppress strong reverberation, the constructed overcomplete dictionary must resemble the features of the target signal as closely as possible while differing from those of the strong reverberation signal. This enables the elimination of the reverberation component during the sparse representation process, leaving only the target signal and thereby achieving both target extraction and reverberation suppression.

For separating target signals from reverberation interference, the fractional Fourier transform (FRFT) has gained attention because it can achieve energy concentration on chirp basis functions, potentially aiding in reverberation suppression. Therefore, many researchers employing FRFT for processing signals in reverberant backgrounds. Yu and Park (2017) used FRFT to obtain Doppler shift, range, and azimuth information of echo signals in strong ocean reverberation. Zhang *et al.* (2014) introduced FRFT into echo signal detection, using a sliding window to segment the echo signal and extracting the start position of the echo when the segmented signal's peak location in the FRFT domain matched that of the transmitted signal. Liang *et al.* (2024) proposed a frequency-domain adaptive matched filter detection method based on FRFT, achieving target detection in reverberation backgrounds through optimal-order transformation and template matching. However, these studies primarily focus on where reverberation energy does not concentrate on the optimal FRFT order. When the reverberation signal shares the same generation mechanism as the target signal, its energy may also concentrate at the optimal order, making it impossible to suppress the reverberation signal using only optimal-order information.

Therefore, considering that FRFT not only effective target features at the optimal order but also represents features on different time-frequency planes at various fractional orders, its multi-order characteristics serve as a powerful basis for distinguishing sonar targets and can be utilized to differentiate between reverberation and target echo signals. Consequently, our study analyzes the energy distribution characteristics of target echoes and reverberation across multiple fractional orders and selects feature sub-domains with strong discriminative power. Subsequently, sparse sub-dictionaries are constructed within each selected sub-domain. An adaptive gradient optimization strategy (Fan, 2024) is employed to train the fusion weights for these multiple domains, achieving feature fusion over multiple fractional orders. The final fused dictionary significantly enhances target focusing and interference suppression capabilities in strong reverberation backgrounds, thereby improving the extraction capability of active sonar target echoes.

2. Sparse representation theory

Sparse representation theory involves representing or approximating a signal using a linear combination of as few atoms as possible from an overcomplete dictionary. Under sparse representation, most of the signal's energy is concentrated in a small number of significant positions, while noise exists in a more dispersed form. Therefore, by extracting a limited number of energy-concentrated feature points, it is possible not only to retain the primary information of the target but also to suppress background noise.

The sparse representation problem is typically modeled as a non-convex optimization problem:

$$\min_x \|x\|_0 \qquad \text{subject to} \qquad y = Dx, \tag{1}$$

where $D = \{g_\gamma\}_{\gamma \in \Gamma}$ is an overcomplete dictionary whose elements g_γ are called atoms, and the number of atoms far exceeds the signal dimension N, and $\|x\|_0$ denotes the l_0 – norm, representing sparsity – the number of non-zero elements in the sparse coefficient vector. The objective is to represent the original signal y using as few atoms as possible. Figure 1 illustrates the principle of sparse representation.

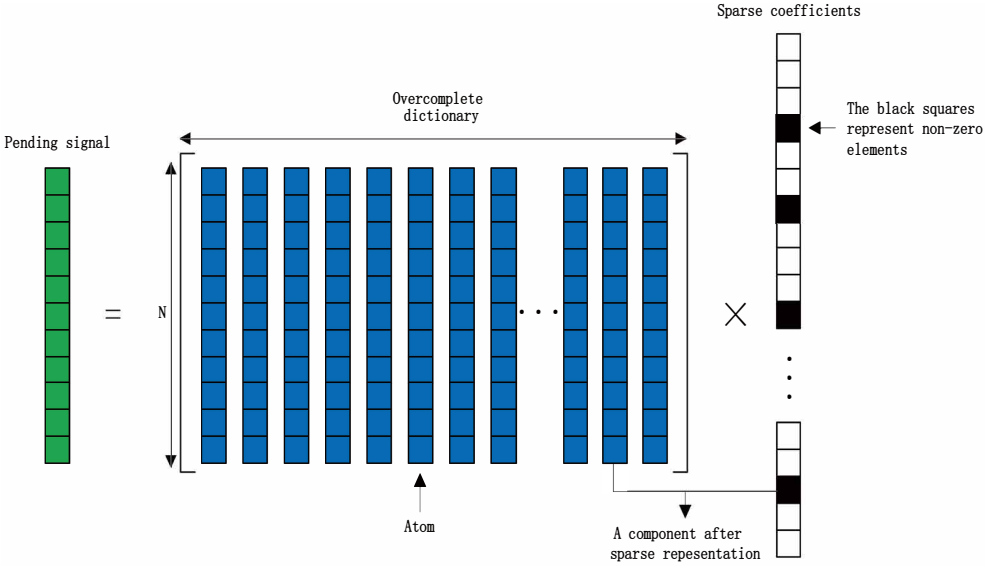

Fig. 1. Schematic diagram of the sparse representation principle.

Sparse representation theory primarily consists of two components: the overcomplete dictionary and the sparse decomposition algorithm. Regarding the construction of the overcomplete dictionary, traditional methods are generally based on time-frequency analysis and are suitable for signals exhibiting sharp local feature variations in the time-frequency domain. If the goal is to represent such signals more sparsely, the corresponding dictionary must encompass richer time-frequency features, thereby introducing redundancy compared to the original time-frequency transform dictionary (Liu *et al.*, 2008).

Let us assume $g(t) = L^2(R)$ is a general window function that satisfies four conditions: (1) it is a continuously differentiable real function, (2) $g(t) \in O\left(1/(t^2+1)\right)$, (3) its norm is 1, i.e., it is equivalent to $\|g\| = 1$, and (4) its integral over the entire real number line is non-zero, i.e., $g(0) = 0$. By applying scaling transformations, time shifts, and modulation to $g(t)$, a generalized overcomplete representation dictionary can be obtained. The characteristics of a well-structured overcomplete dictionary are as follows: first, the dictionary contains a sufficient number of atoms with a wide variety to achieve sparse signal representation and yield high-quality decomposition results, and second, the atoms within the overcomplete dictionary should exhibit significant differences, ensuring storage and computational efficiency. As for the sparse decomposition algorithm, its core lies in iteratively approximating the signal, during which components mismatched with the target structure are eliminated and effective information is retained. For instance, the matching pursuit algorithm selects, in each iteration, the atom from the overcomplete dictionary that exhibits the strongest correlation with the signal to be decomposed.

3. Construction method of the fusion dictionary in the multi-order FRFT domain

The overall workflow of the proposed method is illustrated in Fig. 2, which embodies the core concept of 'multi-order feature extraction + weighted fusion + sparse reconstruction.'

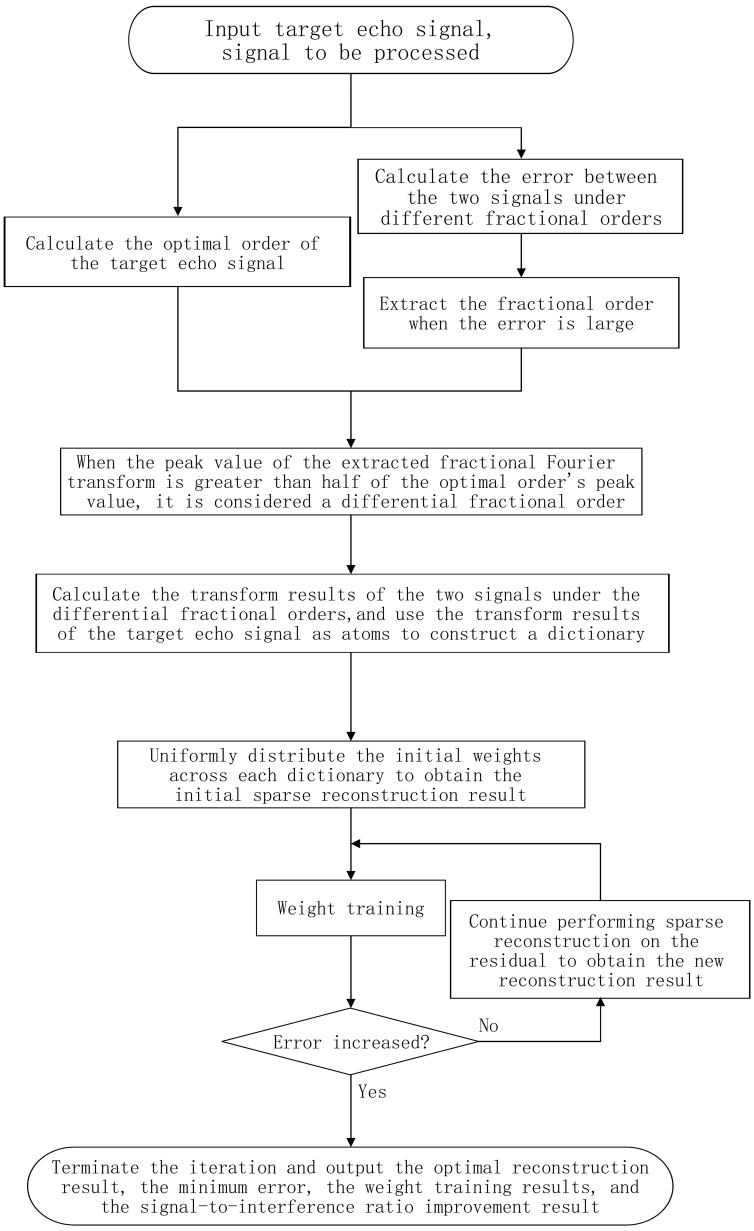

Fig. 2. Flowchart of sparse dictionary construction based on multi-order FRFT domain feature fusion.

The FRFT is a generalized form of the conventional Fourier transform. By introducing a fractional order parameter, it defines a transform domain situated between the time and frequency domains (Ma *et al.*, 2018). Compared to the conventional Fourier transform, which provides only frequency-domain information, the FRFT reveals different features across various time-frequency planes. It offers significant advantages when processing echo signals containing linear frequency modulation (chirp) characteristics. Multi-order FRFT features can capture the distinguishing characteristics between the target signal and strong reverberation (Weng, 2020).

Therefore, the fundamental idea of this paper is as follows: by extracting the differential features between the target echo signal and strong reverberation across different fractional orders, a multi-order sparse dictionary is constructed. Subsequently, an adaptive weighted fusion mechanism is employed to optimize the feature components within each dictionary that contribute most significantly to the target, thereby suppressing the interference signal. Finally, target signal separation and enhancement are achieved through sparse reconstruction.

This method primarily consists of two parts: differential fractional order extraction and fusion dictionary weight training, which are elaborated in Subsec. 3.1 and Subsec. 3.2, respectively.

3.1. Differential fractional order extraction method

In the actual processing of echo signals, since the generation mechanism of strong reverberation is similar to that of the target echo signal, the objective is to identify fractional orders in the FRFT domain that capture the differences between the strong reverberation and the target echo signal. These selected orders are then used to construct a multi-domain dictionary for separating the target echo from the strong reverberation. The specific steps are as follows:

1. Optimal fractional order estimation:
   Apply the FRFT method based on the maximum peak criterion to perform FRFT on the target echo signal. Identify the fractional order at which the signal's energy is most concentrated in the transform domain, denoted as the optimal order. The amplitude at this order serves as the baseline reference for subsequent differential order extraction.
2. Error difference calculation:
   Perform the FRFT on both the target signal and the strong reverberation signal separately across a predefined set of fractional orders. Calculate the error (using the Euclidean distance method) between them at each order to evaluate their structural differences under different fractional orders.
3. Differential fractional order selection:
   Select those fractional orders where the error in the FRFT domain is relatively large and the signals exhibit distinct energy concentration as the differential fractional orders. The selection criterion is that the peak energy at a candidate order is not lower than a specified proportion of the peak energy at the optimal fractional order. These differential orders are considered structurally more suitable for distinguishing the target signal from the interference signal.

Through the such steps, a set of differential fractional orders is obtained. These orders are used to construct the multi-domain sparse dictionary, with each fractional order corresponding to a sub-dictionary within this domain, forming the initial framework for multi-feature fusion sparse representation.

3.2. Weight training method for the fusion dictionary

Suppose P differential fractional orders are extracted. Each order corresponds to a sparse sub-dictionary D_i and is assigned a weight w_i. The construction of each dictionary is based on the FRFT results at that specific order as the atom set, forming dictionaries containing prior feature information. To achieve effective fusion of information from multiple fractional orders, this research designs a weight training method based on gradient descent. By incorporating historical gradient information and neighborhood fluctuation estimation, the learning rate is adaptively adjusted to optimize the contribution of each dictionary to the overall sparse representation.

The specific process is as follows:

1. Weighted model for sparse reconstruction results:
   For the sparse dictionary set constructed from P differential fractional order domains, the overall sparse reconstruction result can be expressed as a weighted superposition of the reconstruction results from each sub-dictionary:
   $$S_{re} = \sum_{i=1}^{P} w_i \varphi_i = \sum_{i=1}^{P} w_i D_i \alpha_i, \tag{2}$$
   where S_{re} is the weighted reconstruction result, φ_i is the reconstruction result corresponding to the i-th dictionary, D_i is the sub-dictionary constructed from the i-th differential fractional order, α_i is the sparse coefficient under that dictionary, and w_i is the corresponding fusion weight. The goal is to train the weights w_i so that the fusion result approximates the original target echo signal S as closely as possible, thereby achieving more accurate reconstruction of the target information.
2. Loss function and gradient calculation:
   To this end, the gradient descent method is introduced for weight training. Aiming to minimize the reconstruction error, the loss function is defined as

$$L = \|S - S_{re}\|_2^2 = \left\|S - \sum_i^P w_i\varphi_i\right\|_2^2 = \|S - \phi w\|_2^2 = (S - \phi w)^T(S - \phi w) = S^T S - S\phi w - w^T\phi^T S + w^T\phi^T\phi w, \quad (3)$$

where S is the target echo signal, $\phi = [\varphi_1, \varphi_2, ..., \varphi_P]$, and $w = [w_1, w_2, ..., w_P]^T$. The gradient corresponding to weight w_i is given by

$$\text{Grad}_i = \partial L/\partial w_i = \sum_{j=1}^P w_j\varphi_i^T\varphi_j + \sum_{j=1}^P w_j\varphi_j^T\varphi_i - S^T\varphi_i - \varphi_i^T S, \qquad i = 1, 2, ..., P. \quad (4)$$

3. Curvature estimation and historical information recording:
To further improve the convergence efficiency of weight updates and enhance the stability of the training process, the second-order derivative of w_i is introduced as $L^{(2)} = 2\varphi_i^T\varphi_i$, which aids in evaluating the curvature of the objective function and informs step size selection. Simultaneously, considering that gradients may fluctuate sharply during actual training, a history-based approach is used to dynamically adjust the learning rate, and update the weights based on past gradient and loss information. The historical maximum absolute gradient is denoted as

$$\text{MaxGrad} = \max\left(\partial L/\partial w_i^{(t)}\right), \quad (5)$$

and the historical maximum absolute loss function value is

$$\text{MaxLoss} = \max\left(\left|L\left(w^{(t)}\right)\right|\right). \quad (6)$$

4. Adaptive neighborhood range adjustment:
To avoid local optima or convergence stagnation due to an excessively small learning rate, an adaptive neighborhood range (Range) is defined, which is gradually adjusted during the iteration process. When approaching the optimum, the local search range for learning samples should be appropriately expanded. The variable Range is obtained as follows:

$$\text{Range} = \log_2(1 + \log_{R_b} R_r), \quad (7)$$

where

$$R_b = 1 + |L'|/\text{MaxGrad} = 1 + \left|\text{Grad}_i'\right|/\text{MaxGrad}, \quad (8)$$

$$R_r = 1 + (|L| + \text{MaxLoss})/\text{MaxLoss}. \quad (9)$$

5. Weight update strategy:
Based on the aforementioned calculations, the main weight update rule is

$$\begin{aligned} w_i &= w_i + LR \times L' = w_i + \frac{\log_{(2+\text{FlucDegree})}(1 + |\text{Grad}_i|)}{|\text{Grad}_i| + \text{FlucDegree}} \times \text{Grad}_i \\ &= w_i + \frac{\log_{(2+\text{Cur}+\text{AdaVar})}(1 + |\text{Grad}_i|)}{|\text{Grad}_i| + \text{FlucDegree}} \times \text{Grad}_i, \end{aligned} \quad (10)$$

where FlucDegree reflects the fluctuation degree of the first-order derivative, and Cur is the current weight. The corresponding computational formulas are:

$$\text{Cur} = \frac{\left|L^{(2)}\right|}{(1 + L'^2)^{3/2}}, \quad (11)$$

$$\text{FlucDegree} = \log_{(2\times\text{Range})}(1 + \text{Var}(\varphi_i)). \quad (12)$$

6. Overall training process:

   a) Initialize weights and sub-dictionaries: initialize weights w_i^0 for all sub-dictionaries generated from differential fractional orders. Perform the first sparse reconstruction to obtain the initial reconstruction result and initial dictionary coefficients.

   b) Alternating update iteration:

      ① fix the sparse coefficients α_i, update the weights w_i,

      ② after updating the weights, perform sparse reconstruction again to obtain new coefficients α_i,

      ③ repeat the prior process until the reconstruction error converges or the preset number of iterations is reached.

   c) Output final results: output the optimal sparse reconstruction result S_{re} and the minimized reconstruction error, which will be used for subsequent signal reconstruction or interference suppression.

Through the aforementioned optimization method, the fusion sparse dictionary, constructed in this study, can adaptively capture the energy characteristics of the target echo signal across different fractional orders, effectively enhancing signal reconstruction quality and interference suppression capability.

4. Experimental results and analysis

4.1. Experimental overview

The data processed in this paper originates from lake trial experiments conducted in the Xin'anjiang waters of Jiande City, Hangzhou, Zhejiang Province. The experimental target was a spherical model. The experimental deployment is shown in Fig. 3. During the experiment, the target moved from far to near or from near to far. A total of 150 test signals were selected for processing. The experiment employed a monostatic active testing system transmitting a linear frequency modulated (LFM) signal with the following parameters: frequency sweep bandwidth of 40 kHz, center frequency of 60 kHz, pulse width of 5 ms, and a sampling frequency of 1 MHz.

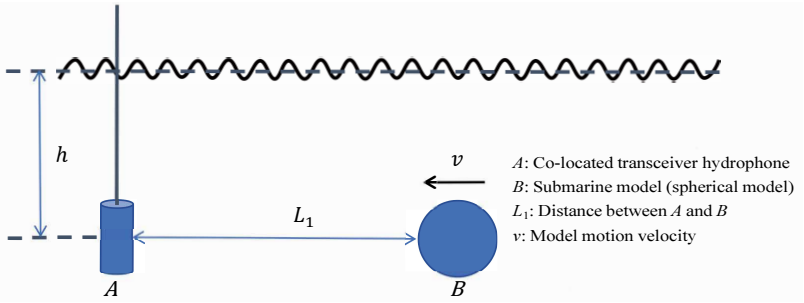

Fig. 3. Experimental deployment.

4.2. Analysis of measured data processing

Based on the test results selected in the previous subsection, where the target's motion was from far to near, the acquired signals were used to construct signals to be processed under different signal-to-reverberation ratios (SRR). The waveform of the 75th test signal under different SRR conditions (3 dB and 5 dB) is shown in Fig. 4.

First, the signal with an SRR of 3 dB was selected as the signal to be processed. To compare the sparse representation performance between the FRFT domain and the time domain, dictionaries were constructed using time-domain and FRFT domain features, respectively. The resulting sparse coefficient plots are shown in Fig. 5.

Based on the sparse coefficient plots, it can be observed that the sparse coefficients for the time-domain dictionary are relatively dispersed, while those for the FRFT domain dictionary are more concentrated. This indicates that the FRFT domain dictionary can better separate the target signal from the reverberation background.

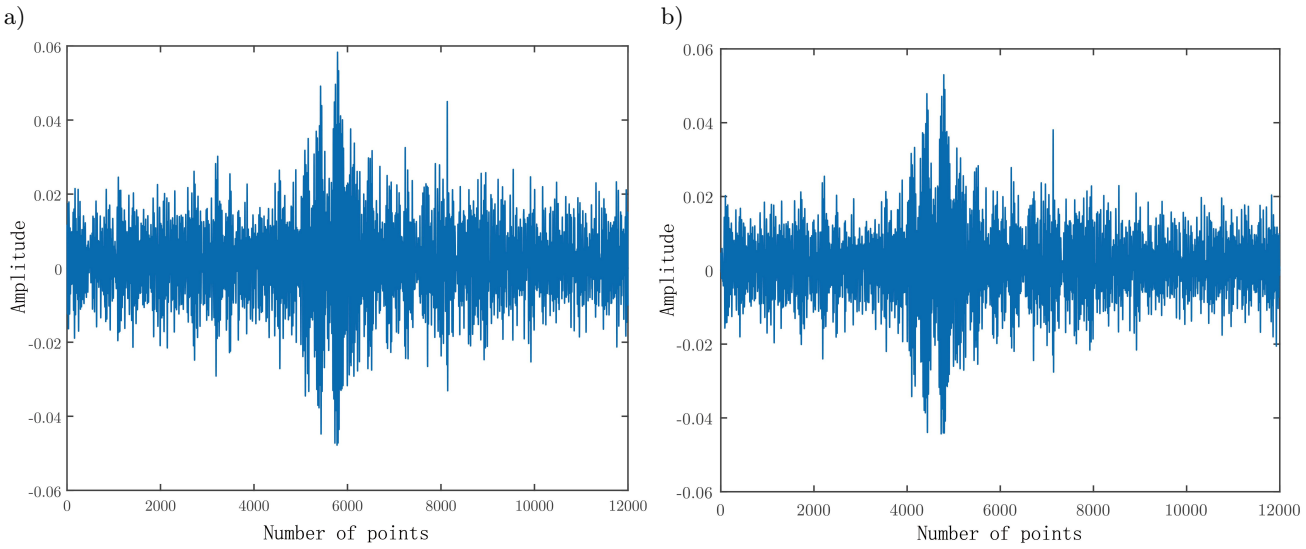

FIG. 4. Waveform of the 75th test signal under different SRR conditions: a) SRR is 3 dB, b) SRR is 5 dB.

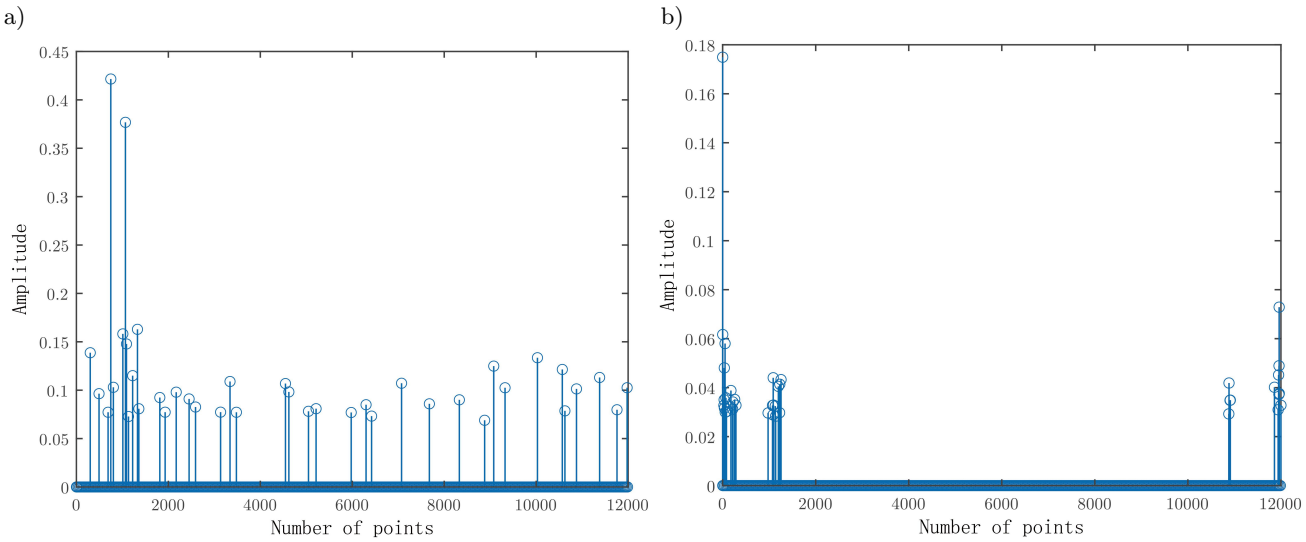

FIG. 5. Comparison of sparse coefficients in time domain vs. FRFT domain:
a) time-domain sparse coefficient plot, b) FRFT domain sparse coefficient plot.

To extract the target echo signal under strong reverberation, the last measurement signal was used as the target signal, and the corresponding reverberation portion from that measurement was extracted to derive the differential fractional orders. Figure 6a shows the amplitude-normalized waveform of the target signal, and Fig. 6b shows the amplitude-normalized waveform of the reverberation signal.

Following the workflow illustrated in Fig. 2, the FRFT was first applied to the target signal to find its optimal fractional order. Using the maximum peak criterion, it was determined that the signal exhibits the strongest energy concentration at a specific fractional order, which is designated as the optimal order for the target signal. The FRFT result of the target echo signal at this optimal order is shown in Fig. 7a, and the corresponding transformation result of the reverberation signal at the same order is shown in Fig. 7b.

When comparing the results of the target echo and reverberation signals at the optimal fractional order in Fig. 7, it can be seen that the reverberation signal also exhibits energy concentration at the target's optimal order, although its concentration amplitude is significantly lower than that of the target signal. Nevertheless, the concentration degree in the reverberation signal is still relatively high. Therefore, it is necessary to identify more discriminative fractional orders for feature-domain processing. According to the proposed method, the differences

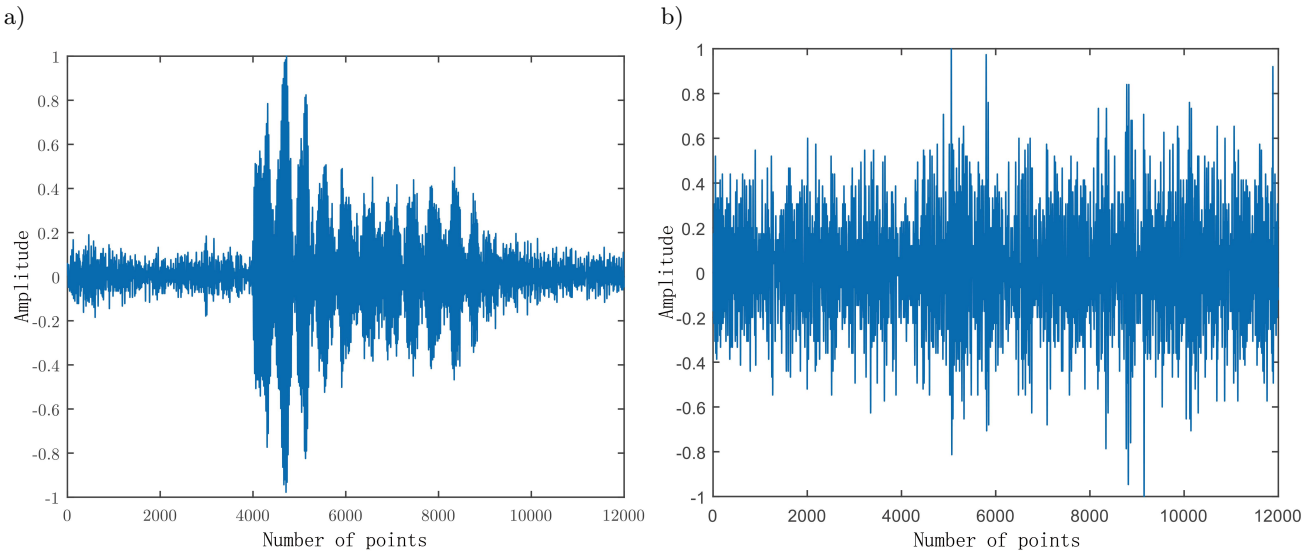

FIG. 6. Amplitude-normalized waveforms of target echo signal (a) and reverberation signal (b).

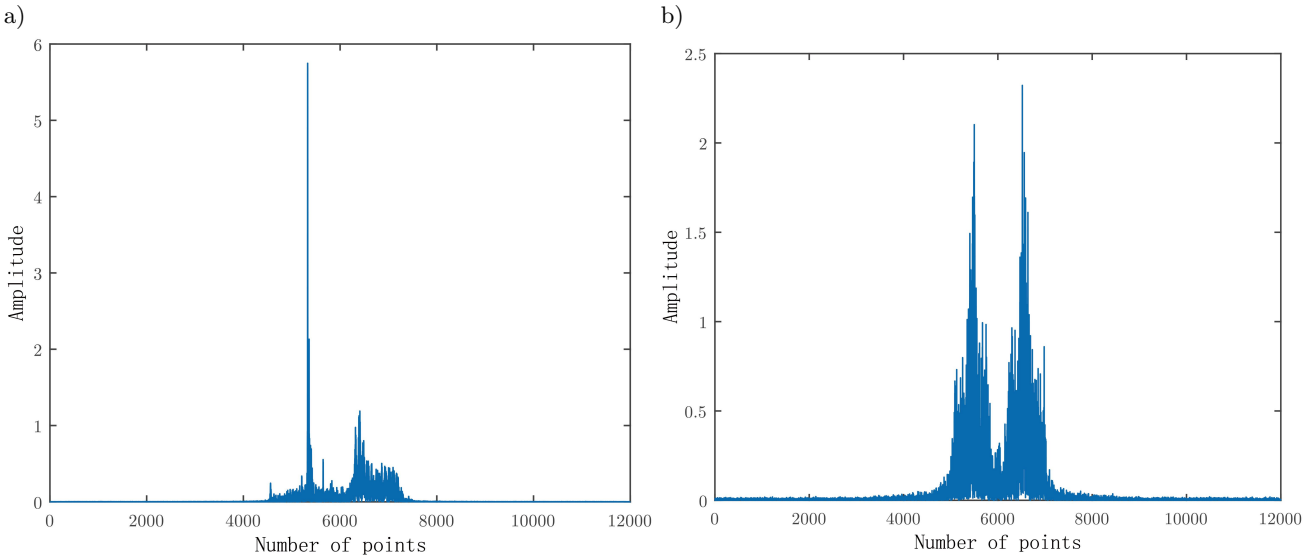

FIG. 7. FRFT results at the optimal fractional order of target signal (a) and reverberation signal (b).

between the target and reverberation signals across different fractional orders were calculated to determine the final set of differential fractional orders, which serve as the basis for constructing the sparse representation dictionary. The calculated differential fractional orders for the target and reverberation signals are:

$$P_{\text{diff}} = [0.939,\ 0.938,\ 0.940,\ 0.937,\ 0.936]\,. \tag{13}$$

Based on the obtained differential fractional orders, the proposed method was used to construct sparse representation dictionaries in each corresponding fractional order domain. The signal to be processed was then analyzed. The final trained weight distribution was $[w_1, w_2, w_3, w_4, w_5] = [0.1642,\ 0.1569,\ 0.1801,\ 0.2324,\ 0.2664]$. The reconstruction result is shown in Fig. 8a, with a corresponding minimum reconstruction error of $\xi_{\min} = 0.68561$. The reconstructed SRR was 18.06 dB, representing an improvement of 15.06 dB. Simultaneously, the proposed method was applied to process the 75th test signal with an initial SRR of 5 dB, and the processing result is shown in Fig. 8b. By comparing the signals before and after reconstruction, it is evident that the target signal, originally submerged in reverberation, has been effectively extracted.

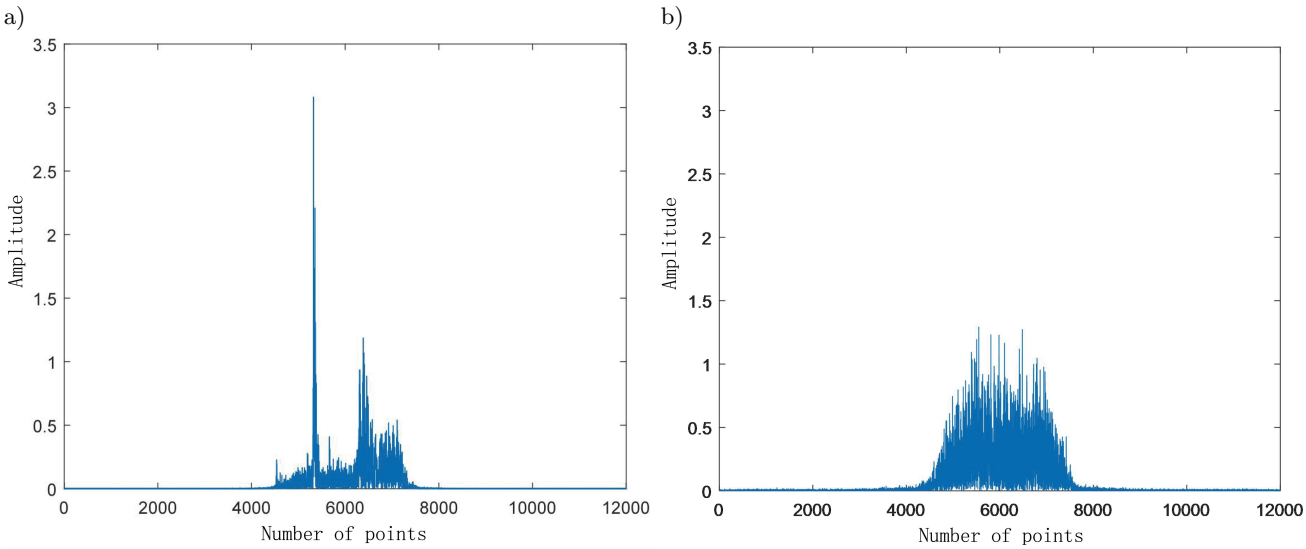

FIG. 8. Sparse reconstruction results (of the 75th test signal) under different initial SRR conditions for SRR = 3 dB (a) and SRR = 5 dB (b).

Figure 9 shows the SRR improvement for the 75th test signal across different initial SRR conditions. It demonstrates that the proposed method maintains high target signal extraction performance even in strong reverberation backgrounds.

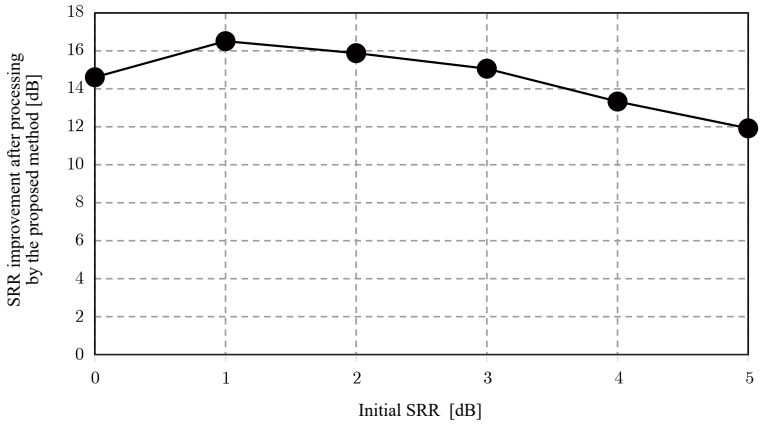

FIG. 9. SRR improvement under different initial SRR conditions.

The SRR improvement for different test signals under various initial SRR conditions were obtained, and the data are presented in Table 1.

TABLE 1. SRR improvement [dB] for different test signals under different initial SRR conditions.

Test signals	Initial SRR					
	0	1	2	3	4	5
25th	6.53	12.17	10.97	9.74	9.08	8.15
55th	13.96	13.88	12.67	11.27	10.24	9.18
110th	10.48	9.58	8.71	7.67	10.28	9.27
140th	7.46	6.52	12.46	11.60	10.64	9.77

Based on the data in Table 1, it can be concluded that the post-processing SRR improvement is influenced not only by the initial SRR but also by the target's position. When the target is far away and interference is strong, extracting the target signal is more difficult. Furthermore, since measured reverberation was used to construct signals with different initial SRR levels, the actual SRR is lower than the values calculated in

this paper, which also affects the observed improvement. However, the data show that the proposed method achieves effective improvement within the selected SRR range (0 dB to 5 dB), with an improvement of no less than 5 dB, confirming the robustness of the method.

To further validate the method's performance across all test signals, processing was conducted on all test signals under SRR conditions of 3 dB and 5 dB. The SRR improvement for different test signals under these conditions is shown in Fig. 10.

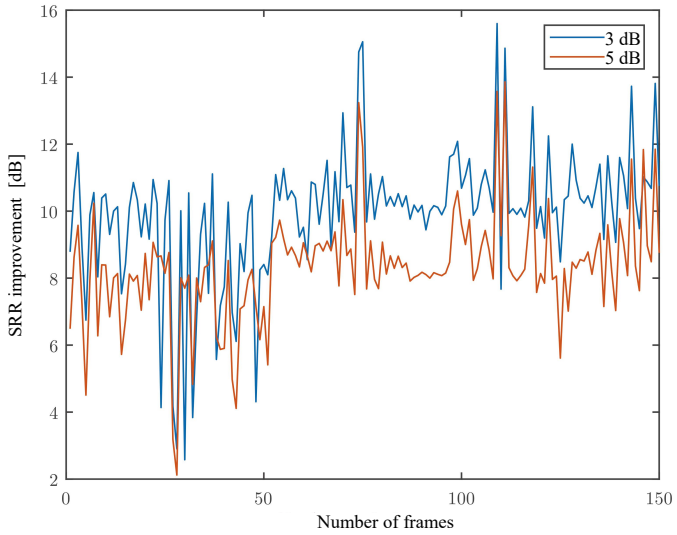

Fig. 10. SRR improvement across all test signals under initial SRR conditions of 3 dB and 5 dB.

Based on the aforementioned processing results, it is concluded that the proposed method successfully implements sparse dictionary construction based on multi-order FRFT domain feature fusion. Experimental results verify the feasibility of the method. The achieved SRR improvement is no less than 2 dB. The method enables effective extraction of target echoes and suppression of reverberation in strong reverberation backgrounds, demonstrating good practical application prospects.

5. Conclusion

This paper addressed the challenge of strong reverberation interference suppression in underwater acoustic signal processing by proposing a sparse dictionary construction method based on multi-order FRFT domain feature fusion. Centered on the core concept of 'multi-order extraction + weighted fusion + sparse reconstruction,' the method achieves effective extraction of target echo signals and suppression of interference. Processing and analysis of measured data demonstrated that the target echo in the output signal was effectively enhanced. Under conditions of low initial SRR ranging from 0 dB to 5 dB, the method consistently achieved an SRR improvement of no less than 2.1 dB, with a maximum improvement of 15.6 dB. These results fully illustrate the method's capability to enhance SRR and suppress interference in strong reverberation backgrounds, as well as its adaptability and robustness to complex underwater acoustic environments. The proposed method not only theoretically extends the application boundaries of the FRFT within sparse representation but also provides an effective technical approach for addressing strong interference suppression issues in underwater acoustic detection and recognition systems.

Fundings

This work was supported by the Joint National Natural Science Foundation of China (no. U22A2044) and the Key Laboratory Fund from Underwater Test and Control Technology (no. 2023-JCJQ-LB-030).

Conflict of interest

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Authors' contribution

Tongjing Sun: data curation, formal analysis, investigation, resources, writing – review and editing, funding acquisition, supervision. Lei Chen: conceptualization, data curation, investigation, methodology, writing – original draft. Xiaohong Deng: data curation, formal analysis, writing – review and editing. All authors reviewed and approved the final manuscript.

Acknowledgments

This article uses experimental data collected at Shanghai Jiaotong University (China), and we gratefully acknowledge our colleagues for their experimental skills.

References

1. Cui X., Chi C., Li S., Li Z., Li Y., Huang H. (2023), Waveform design using coprime frequency-modulated pulse trains for reverberation suppression of active sonar, *Journal of Marine Science and Engineering*, **11**(1): 28, https://doi.org/10.3390/jmse11010028.
2. Deng B., Tao R., Qi L., Liu F. (2005), A study on anti-reverberation method based on fractional Fourier transform, *Acta Armamentarii*, **26**(6): 761–765, https://doi.org/10.3321/j.issn:1000-1093.2005.06.010.
3. Fan Y. (2024), *Parameter adaptive setting and global optimization improvement for gradient descent methods*, Master's thesis, Northwest Normal University.
4. Ge F., Zhang Y. (2023), Review of weak underwater acoustic signal processing, *Journal of Signal Processing*, **39**(10): 1728–1747, https://doi.org/10.16798/j.issn.1003-0530.2023.10.002.
5. Lei L. (2014), *Sparse representation of signals via wavelet modulus maxima and compressed sensing reconstruction*, Master's thesis, Beijing Jiaotong University.
6. Liang Q., Hu P., Chen Z., Li J. (2024), Frequency domain matching filter detection method based on FRFT, *Ship Science and Technology*, **46**(05): 69–73.
7. Liao M., Zhang X., Zhang X. (2014), Classification and recognition of underwater acoustic signal based on sparse representation, *Journal of Detection & Control*, **36**(04): 67–70+77.
8. Liu D., Shi G., Zhou J. (2008), New method for signal sparse decomposition over a redundant dictionary, *Journal of Xidian University*, pp. 228–232.
9. Liu G., Zhou X. (2018), Interfere suppression and communication signal reconstruction method based on signal sparse representation in complex electromagnetic environment, *Computer Systems & Applications*, **27**(11): 149–154, https://doi.org/10.15888/j.cnki.csa.006613.
10. Ma J., Miao H., Su X., Gao C., Kang X., Tao R. (2018), Research progress in theories and applications of the fractional Fourier transform, *Opto-Electronic Engineering*, **45**(06): 5–28, https://doi.org/10.12086/oee.2018.170747.
11. Meng W. (2024), *Research method of underwater acoustic direction of arrival estimation based on noise integral sparse Bayesian learning*, Master's thesis, South China University of Technology.
12. Sun T., He J., Gu Y. (2016), Underwater echo signal processing method based on sparse decomposition, *Journal of Data Acquisition and Processing*, **31**(02): 282–288, https://doi.org/10.16337/j.1004-9037.2016.02.007.
13. Wang H. (2020), *Research on active sonar target classification and recognition based on sparse representation*, Master's thesis, Hangzhou Dianzi University.
14. Weng Y. (2020), Detection and sorting algorithm of multi-component LFM signals based on fractional Fourier transform, *Aerospace Electronic Warfare*, **36**(02): 33–37+55, https://doi.org/10.16328/j.htdz8511.2020.02.009.

15. Wu Y., Xing C., Zhang D., Xie L. (2021), A denoising processing method for the low-frequency underwater acoustic signal based on K-SVD, *Journal of Yunnan University of Nationalities (Natural Sciences Edition)*, **30**(04): 387–393.

16. Yang J. (2023), The application development of sonar technology in marine resource exploration, *Audio Engineering*, **47**(12): 27–29, https://doi.org/10.16311/j.audioe.2023.12.006.

17. Yu G., Park S. (2017), Multipe targets detection in strong reverberation environments, *Journal of Huazhong University of Science and Technology (Natural Science Edition)*, **45**(03): 89–93, https://doi.org/10.13245/j.hust.170316.

18. Zhang Z., Yang X., Dai W. (2014), Pick-up of echo starting location in reverberation, *Ship Electronic Engineering*, **34**(11): 73–75+79.

19. Zhou Z., Wu J., Yuan C., Bai M., Gui Z. (2023), A novel K-SVD dictionary learning approach for seismic data denoising, *Oil Geophysical Prospecting*, **58**(05): 1072–1083, https://doi.org/10.13810/j.cnki.issn.1000-7210.2023.05.005.

Archives of Acoustics Vol. 51, No. 2, pp. 249–266 (2026), https://doi.org/10.24423/archacoust.2026.4315

Research Paper

Few-Shot Transfer for Speech Enhancement Using SEGAN with Stability Guardrails

Rubi Sharma*, Firos A.

Department of Computer science & Engineering, Rajiv Gandhi University
Arunachal Pradesh, India

*Corresponding Author: rubi.sharma@rgu.ac.in

Received September 1, 2025; revised February 15, 2026; accepted February 18, 2026; available online March 10, 2026; version of record June 16, 2026; published issue June 24, 2026.

High-quality speech communication is often compromised by background noise, reducing intelligibility and perceived quality. We investigate data-efficient few-shot transfer of the speech enhancement generative adversarial network (SEGAN) to a new noise domain. Starting from a generator pre-trained on VoiceBank–DEMAND, we adapt the model to MiniLibriMix using only 300 paired noisy–clean examples. To prevent overfitting and catastrophic forgetting, we introduce stable adversarial few-shot enhancement (SAFE), a three-fold stabilisation strategy with (1) exponential-moving-average (EMA) weight averaging, (2) L2-SP weight anchoring to the source-domain parameters, and (3) a teacher–student consistency loss. SAFE maintains VoiceBank performance (PESQ ≈ 1.84; STOI ≈ 90 %) and, after an optional perceptual fine-tuning stage (L_1 + MR-STFT), yields substantial target-domain gains on MiniLibriMix (PESQ 1.11 → 1.26, STOI 71.5 % → 81.5 %) with only a minor source-domain trade-off in STOI. Ablation experiments demonstrate that EMA provides the strongest stabilising effect, while L2-SP and consistency regularisation offer complementary benefits. These results suggest that stable few-shot adaptation may make lightweight time-domain speech enhancers practical for rapid deployment in novel acoustic environments.

Keywords: speech enhancement, generative adversarial networks, few-shot learning, transfer learning, domain adaptation, stability regularization.

1. Introduction

Deep learning has greatly advanced single-channel speech enhancement (SE) in recent years, with models increasingly able to remove noise and improve speech quality. Most state-of-the-art SE systems are data-hungry, requiring large corpora of paired noisy and clean speech for training. For example, modern architectures like MetricGAN variants and diffusion models can achieve high perceptual quality (PESQ > 3.0) and intelligibility (STOI >0.94) on benchmark datasets, but these models are generally trained from scratch on training sets comprising dozens of hours. In practical scenarios, however, one often needs to deploy SE models in a new domain (e.g., a different noise profile or language) when we have very little labeled data. Under such domain shift conditions, a model trained on one dataset may perform poorly on another due to mismatched noise characteristics or speaker differences. This underscores the need for few-shot transfer learning techniques to efficiently adapt SE models to new domains using minimal data.

Generative adversarial networks have been a popular approach for SE, starting with the speech enhancement generative adversarial network (SEGAN) by Pascual *et al.* (2017). SEGAN introduced a waveform-to-waveform enhancement model (the convolutional neural networks (CNN) autoencoder with skip connections) trained with a GAN objective, and demonstrated notable improvements in perceptual quality on the VoiceBank + DEMAND dataset. Follow-up works explored adapting SEGAN to new conditions. Pascual *et al.* (2017) fine-tuned SEGAN for new languages and noise types, showing that with as little as 10 min of target data the

model could approach the performance of full-data training. Hou *et al.* (2019) proposed a domain-adversarial training strategy to make SEGAN's features noise-invariant, improving generalization to unseen noise without requiring extensive target data. Other researchers have explored architectural modifications to make SEGAN more adaptable: Li *et al.* (2021) introduced sinc-SEGAN, replacing the first convolutional layer with a parameterized sinc filter to better capture speech bandwidth, which eased fine-tuning and reduced the model size. Recently, Lv *et al.* (2024) combined self-attention and temporal convolutional networks in SASEGAN-TCN, achieving higher base performance and showing improved noise suppression under unseen conditions. Multi-task and cross-domain transfer approaches have also been investigated; for example, Wang *et al.* (2020) leveraged automatic speech recognition (ASR) task knowledge by using paired senone classifiers to guide SEGAN adaptation to new noise types. In parallel, knowledge distillation and test-time adaptation techniques have emerged: Kim and Kim (2021) used zero-shot learning with knowledge distillation for personalized speech enhancement. These related works underscore the variety of transfer learning strategies for SE, ranging from simple fine-tuning and feature reuse to adversarial domain adaptation, meta-learning, and knowledge distillation.

We present a study on few-shot domain adaptation for speech enhancement, highlighting methods that enable efficient model adaptation under low-resource conditions. We assume a high-performance SE model is available for a well-resourced source domain, and we have only a handful of noisy-clean pairs (on the order of a few minutes of speech) for a low-resource target domain. Our goal is to adapt the model to perform well on the target domain while preserving its performance on the source domain (i.e., avoiding catastrophic forgetting). We choose SEGAN as the base model due to its proven efficacy on VoiceBank–DEMAND and its relatively lightweight architecture that can be fine-tuned quickly. To achieve stable adaptation on limited data, we introduce a SAFE adaptation strategy – few-shot transfer with stability guardrails – which incorporates several regularization and consistency techniques into the fine-tuning process. The key contributions of our work include: (1) demonstrating successful few-shot transfer of a time-domain SEGAN model to a new domain (from environmental noise to two-speaker mixture 'noise') with only 300 training examples, (2) proposing a combination of EMA-based weight averaging, L2-SP weight regularization, teacher-student consistency, and source-target data mixing to stabilize few-shot adaptation on limited samples, and (3) providing a detailed analysis of the impact of each stabilizing technique via ablation studies. We report objective speech quality and intelligibility metrics, including PESQ and STOI on both the source domain and target test sets to confirm that model adaptation improves target-domain performance without degrading overall quality. To the best of our knowledge, this is the first application of mean-teacher consistency and L2-SP regularization in combination for SEGAN domain adaptation. While our adapted SEGAN does not seek to exceed the absolute performance of larger, modern architectures (e.g., transformer- or diffusion-based models), it provides a data-efficient transfer framework that could be extended to such models in future research. Furthermore, we highlight how our approach complements existing transfer learning strategies and outline future directions, including applying SAFE to state-of-the-art SE backbones and exploring the unsupervised domain adaptation.

The remainder of this paper is organized as follows: Section 2 reviews related work on transfer learning approaches in speech enhancement. Section 3 outlines the methodology, comprising the model architecture and adaptation procedures. Section 4 describes the experimental setup. Section 5 reports the results and ablation studies. Section 6 discusses the results, Sec. 7 presents the significance and benefits, and Sec. 8 concludes with directions for future work.

2. Related work

2.1. Transfer learning in speech enhancement

Transfer learning has been explored in SE to handle domain mismatch and low-resource scenarios. A straightforward approach is fine-tuning a pretrained model on new data – Pascual *et al.* (2018) showed that SEGAN can be fine-tuned on a new language or noise condition with a small dataset, achieving performance comparable to training from scratch with much more data. However, naive fine-tuning may overfit when only a few examples are available. To address this, researchers have developed methods to leverage unpaired or unlabeled data from the

target domain. Domain adaptation techniques often employ adversarial objectives: Liao *et al.* (2019) introduced a domain-adversarial training where a discriminator forces the SE model's encoded features to be indistinguishable between seen and unseen noise domains, yielding robust enhancement on non-stationary noises. Hou *et al.* (2019) similarly used a domain classifier to guide SEGAN to learn noise-invariant representations, improving generalization to unseen DNS challenge noises (Reddy *et al.*, 2020). Another line of work is meta-learning for SE: for example, Yu *et al.* (2021) proposed OSSEM, a one-shot speaker adaptive SE method that uses meta-learning to adapt a pretrained SE model to a particular speaker using only a single utterance. Their approach demonstrates that meta-training on multiple tasks enables rapid adaptation and improved performance on unseen speakers; however, this meta-training phase can be computationally intensive. This concept is akin to few-shot learning and has shown promise in other speech tasks, but is not yet widely adopted in SE due to complexity.

2.2. Architectural and training improvements

Several works modify the SE model architecture to facilitate transfer. The sinc-SEGAN model by Li *et al.* (2021) used sinc convolutional filters to hard-code prior knowledge of speech bandwidth, which allowed the model to train faster and maintain performance even after removing some encoder layers (reducing model size). This kind of inductive bias can be seen as a form of transfer learning, since a model with fewer parameters is less prone to overfitting on new small data. Self-attention mechanisms, such as those employed in SASEGAN-TCN by Lv *et al.* (2024), have been integrated to enhance the representational capacity of the baseline SEGAN. While these improved architectures achieve better base performance, our work is orthogonal in that we focus on stabilizing the training process for domain adaptation rather than proposing a new architecture. Table 1 summarizes existing SEGAN-based transfer strategies, highlighting their respective datasets and adaptation mechanisms.

Table 1. Comparison of SEGAN transfer learning and related works.

Research	Architecture	Transfer strategy / dataset(s)	Key contribution	Performance gain
Pascual *et al.* (2018)	Original SEGAN (U-Net + GAN)	Inter-language and noise transfer on VoiceBank + DEMAND	Demonstrated SEGAN can be fine-tuned efficiently on new languages/noises	Improved PESQ/STOI on new domains
Hou *et al.* (2019)	SEGAN + domain classifier	Domain-adversarial training using DNS challenge and CHiME (Barker *et al.*, 2015)	Learned noise-invariant features via adversarial training	Strong generalization to unseen noise
Li *et al.* (2021)	SEGAN + sinc convolutions	Lightweight CNN, pretrained encoder reuse (VoiceBank)	Lower complexity and easier fine-tuning with fewer parameters	Maintained SEGAN performance with fewer parameters
Lv *et al.* (2024)	SEGAN with self-attention and TCN	Pretrained SEGAN enhanced by self-attention and temporal convolution (DNS challenge)	Improved temporal modeling and noise suppression	Significant STOI/PESQ improvement
Vinotha *et al.* (2024)	SepFormer with hierarchical attention	Multi-stage transfer learning for dysarthric speech	Improved clarity	Outperformed SEGAN
Wang *et al.* (2020)	SEGAN + senone classifier	Cross-task transfer combining SE and ASR (VoiceBank + CHiME)	Joint enhancement–ASR adaptation via paired senone classifiers	Boosted ASR accuracy under noise
Liao *et al.* (2018)	SEGAN + domain-adaptation layers	Few-shot noise adaptation via adversarial training (DNS)	Robust to unseen noise with limited data	Robust performance in novel environments

3. Methodology

3.1. Baseline SEGAN architecture

Although the study is motivated by SEGAN-style waveform-to-waveform enhancement, the archived implementation uses a compact residual 1D U-Net generator (see Fig. 1). The generator processes a noisy waveform x and predicts a residual noise component $G(x)$; the enhanced waveform is obtained as $\widehat{y} = x - G(x)$.

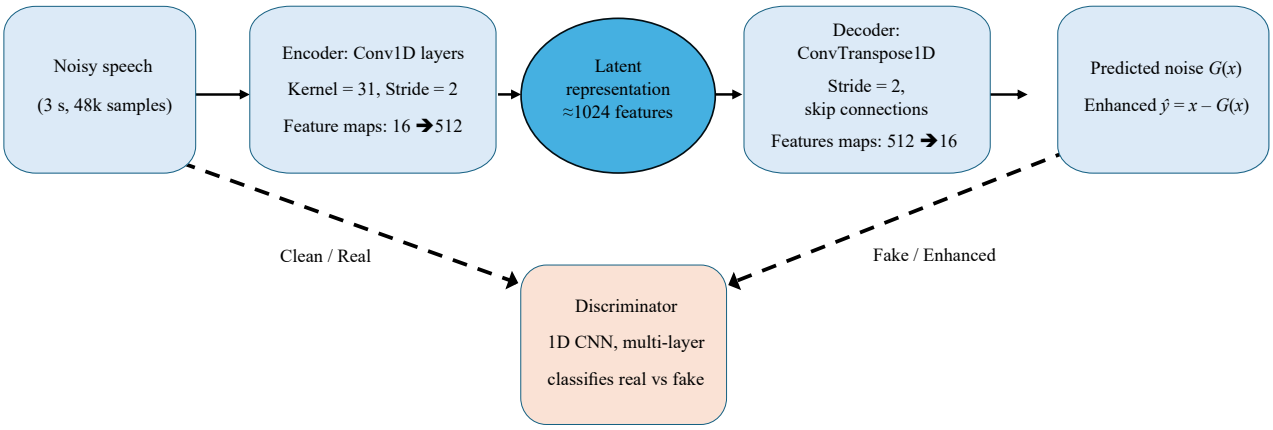

FIG. 1. Residual waveform enhancement architecture used in the reported implementation.

The implemented generator contains four encoder blocks, one bottleneck block, and four decoder blocks. Each encoder block uses a one-dimensional convolution followed by Leaky ReLU activation and instance normalization, with average-pooling downsampling between encoder levels. The decoder uses nearest-neighbour upsampling followed by ConvTranspose1D blocks with ReLU activation and instance normalization. Skip connections concatenate decoder activations with the corresponding encoder features. A final 1×1 convolution maps the decoder output to the residual waveform.

The archived implementation used for the reported experiments does not include a stochastic latent vector or a separately optimized discriminator. Accordingly, the reported training and fine-tuning stages optimize the generator-side waveform enhancement model with reconstruction, regularization, consistency, and MR-STFT losses, as described further.

3.2. Transfer learning strategies

Figure 2 illustrates the four transfer-learning approaches employed with SEGAN:

1. Fine-tuning: retraining the entire model on new noisy datasets for improved generalization.
2. Few-shot transfer: training with limited labeled samples from the target domain to minimize computational overhead.
3. Domain adaptation: adding domain classifiers and aligning latent features to bridge distribution gaps.
4. Knowledge distillation: compressing the large SEGAN model into a smaller student model for efficiency.

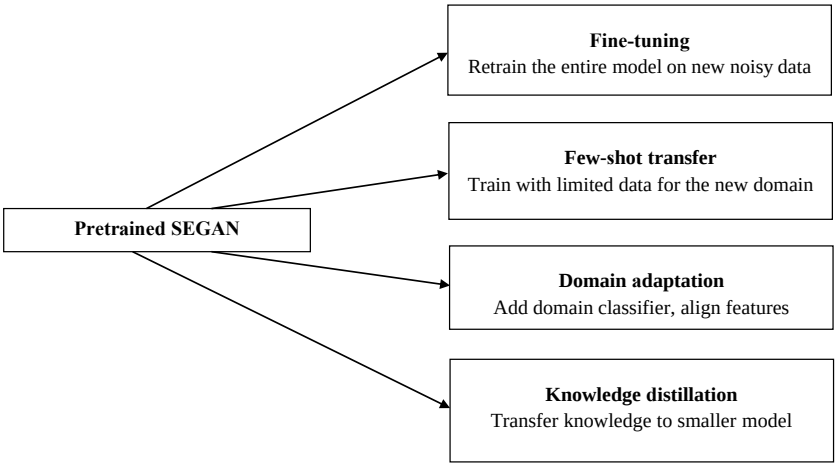

FIG. 2. Transfer learning strategies for SEGAN.

Transfer learning paradigms applied to SE are contrasted in Table 2, outlining their mechanisms and trade-offs.

Table 2. Comparison of transfer learning approaches for speech enhancement.

Approach	How it works	Pros	Cons	Example in SE
Fine-tuning	Retrain the pretrained model on new data (all layers)	Simple, improves adaptation	Needs more data, risk of overfitting	SEGAN retrained on CHiME noise
Few-shot transfer	Train only part of the model (e.g., decoder) with very little new data	Works with few samples, fast training	May not adapt perfectly if noise is very different	Our proposed few-shot SEGAN
Feature extraction	Use pretrained model to extract features, train a new small model on them	Very fast, minimal training needed	Limited improvement, may not capture new noise patterns	Using SEGAN encoder as feature extractor
Domain adaptation	Add domain classifier to learn noise-invariant features	Great for unseen environments	Needs some target domain data, training is complex	Hou *et al.* (2019) (domain-adversarial SEGAN)
Multi-task learning	Train one model on multiple related tasks at once (e.g., SE + ASR)	Learns generalizable features, improves performance	Harder to train, needs multiple datasets	SEGAN + ASR enhancement pipeline
Meta-learning	Model learns to adapt quickly to new tasks with minimal updates (MAML)	Excellent for few-shot cases, very adaptive	Requires complex setup and meta-training	Experimental meta-learning SEGAN (not widely used yet)
Knowledge distillation	A large teacher model trains a small student to mimic it	Creates lightweight models, great for mobile deployment	May lose some performance compared to the teacher	SEGAN-Lite distilled from full SEGAN

3.3. Few-shot SAFE adaptation (stage 1)

We adapt the pretrained SEGAN generator to the MiniLibriMix target domain using stable adversarial few-shot enhancement (SAFE). The generator is initialized with pretrained weights from VoiceBank training. During adaptation, most layers are constrained or frozen to prevent overfitting. Specifically, we set unfreeze_last_k = 1, so that only the final decoder block is directly trainable, while the rest are strongly regularized by weight constraints. The three components that stabilize the adaptation are described further.

3.3.1. Exponential moving average (EMA)

During training, we maintain the EMA of the model parameters, defined as

$$\widetilde{\theta}_t = \alpha\,\widetilde{\theta}_{t-1} + (1-\alpha)\theta_t, \tag{1}$$

where α denotes the EMA decay factor (set to $\alpha = 0.995$ per training step). The EMA parameters $\widetilde{\theta}_t$ produce a temporally smoothed version of the generator and are used for parameter averaging and final evaluation smoothing. EMA is known to stabilize training and improve generalization in semi-supervised learning; here it serves as a buffer against the noisy gradient updates from very limited data.

3.3.2. L2-SP regularization

We apply L2-starting-point (L2-SP) regularization to the generator's weights by adding a penalty term:

$$L_{\text{L2-SP}} = \lambda_{\text{sp}} \|\theta_{\text{adapt}} - \theta_{\text{base}}\|^2, \tag{2}$$

where θ_{base} is the pretrained weight from the source domain and θ_{adapt} is the current fine-tuned weight. A small value of λ_{sp} (set to 1×10^{-4} as in (Li *et al.*, 2018)) encourages the adapted parameters to remain close to their source initialization, thereby reducing catastrophic forgetting of knowledge acquired from VoiceBank. Unlike standard L2 regularization, which penalizes deviation from zero, L2-SP penalizes deviation from the specific pretrained parameter values, acting as a model reuse prior.

3.3.3. Teacher-student consistency

Along with weight-space regularization, we impose an output-space consistency constraint using a frozen teacher model. The teacher is a frozen copy of the pretrained source-domain generator, while EMA is maintained separately for parameter averaging and evaluation smoothing. On source-replay minibatches, the frozen teacher produces a reference enhanced output $\widetilde{y}$ for a given noisy input, and the current generator (student) with parameters θ produces the output $\widehat{y}$. A consistency loss

$$L_{\text{cons}} = \lambda_c \|\widehat{y} - \widetilde{y}\|_1 \tag{3}$$

is added to enforce the student's output to remain close to the teacher's output. This discourages the fine-tuned model from deviating excessively from the stabilized teacher model. The approach is inspired by the mean-teacher paradigm (Tarvainen, Valpola, 2017) in semi-supervised learning, where a student network learns from a temporal average of itself. In our case, because the teacher is a frozen copy of the pretrained source-domain generator, L_{cons} directly promotes consistency with the source-initialized model's behaviour during few-shot adaptation. We set $\lambda_c = 0.1$ based on preliminary experiments.

These constraints are applied concurrently during few-shot fine-tuning. Importantly, no discriminator update is performed in this stage; the adaptation objective uses reconstruction, L2-SP regularization, and consistency losses to prevent divergence under limited data conditions. The overall adaptation loss is therefore defined as

$$L_{\text{adapt}} = L_{\text{rec}} + L_{\text{L2-SP}} + L_{\text{cons}}, \tag{4}$$

where L_{rec} denotes the reconstruction loss, $L_{\text{L2-SP}}$ the L2-SP regularization term, and L_{cons} the teacher–student consistency loss.

3.3.4. Reconstruction loss

For L_{rec}, we use the standard L_1 loss computed between the enhanced waveform $\widehat{y}$ and the clean target waveform y:

$$L_{\text{rec}} = \|\widehat{y} - y\|_1. \tag{5}$$

The L_1 loss (mean absolute error, MAE) is chosen over L_2 (mean squared error) because it generally produces fewer artifacts in the enhanced speech. During few-shot adaptation, L_{rec} serves as the primary driving loss since the adversarial (GAN) loss is disabled. We also experimented with adding a small spectral magnitude loss on the short-time Fourier transform (STFT) of the enhanced and clean signals but observed negligible improvement; therefore, it is omitted in the main training for simplicity. During full SAFE adaptation, minibatches use a 2:1 VoiceBank-to-MiniLibriMix replay ratio to retain source generalization while adapting to the new domain. Table 3 summarizes the symbols and hyperparameters used in the SAFE formulation. Figure 3 illustrates the SAFE adaptation architecture.

Table 3. Symbols and hyperparameters used in SAFE.

Symbol	Definition	Value
θ	Generator (student) parameters	–
θ_0	Pretrained generator parameters	–
$\widetilde{\boldsymbol{\theta}}_t$	EMA-averaged parameters at iteration t	–
x, y	Noisy input; clean target	–
$\widehat{y}$, $\widetilde{y}$	Student output; teacher output	–
α	EMA decay (Eq. (1))	0.995
λ_{sp}	L2-SP weight (Eq. (2))	1×10^{-4}
λ_c	Consistency weight (Eq. (3))	0.1
L_{rec}	L_1 reconstruction $\|\widehat{y} - y\|_1$	–
N	Samples per segment	48 000 (3 s @ 16 kHz)

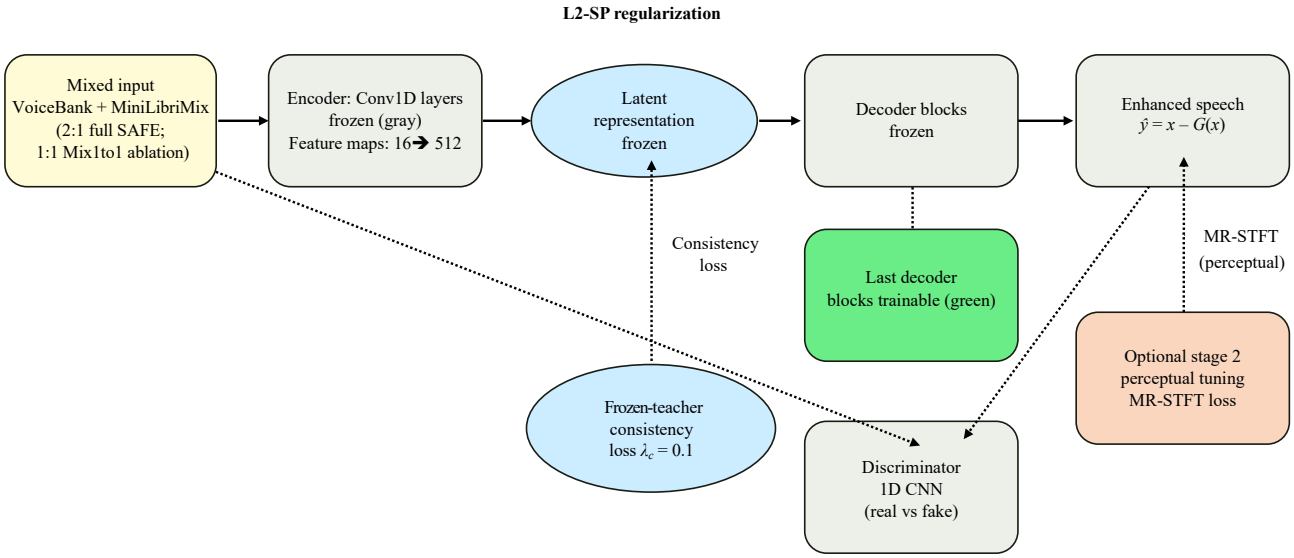

Fig. 3. SAFE adaptation architecture.

3.4. Perceptual fine-tuning (optional; stage 2)

3.4.1. Partial freezing

In practice, the entire generator is allowed to update during fine-tuning, but the L2-SP constraint strongly limits changes in the earlier layers. Alternatively, most layers could be frozen while fine-tuning only the final few layers. In our hyperparameter configuration, we set unfreeze_last_k = 1, so that initially only the final layer was trainable. However, with L2-SP regularization applied, updating all layers proved acceptable because the regularizer naturally constrained the earlier layers to remain close to their pretrained values. Consequently, our final configuration used all layers as trainable, albeit with a very small learning rate for most layers to prevent large deviations.

After the constrained few-shot adaptation stage, the generator generalizes moderately well to the target domain while retaining source-domain performance. We then optionally perform a second fine-tuning stage focused on enhancing perceptual quality on the target domain. In this stage, the generator is further optimized with L_1 waveform reconstruction and multi-resolution short-time Fourier transform (MR-STFT) loss (Shi *et al.*, 2023) to sharpen the output speech.

The MR-STFT loss is computed by applying the short-time Fourier transform to both the enhanced and clean signals using FFT/window sizes of 256, 512, and 1024 samples and summing the L_1 differences between magnitude spectra across these resolutions. This loss emphasizes spectral details at different time–frequency scales and correlates with perceptual audio quality. The overall loss for this stage is defined as

$$L_{\text{perc}} = L_1(\widehat{y}, y) + 0.5 L_{\text{MR-STFT}}(\widehat{y}, y), \tag{6}$$

where the MR-STFT term is weighted by 0.5 and is computed at three resolutions (Yamamoto *et al.*, 2020). No discriminator update or adversarial optimization is performed during this stage.

Unlike the few-shot adaptation stage, L2-SP and teacher–student consistency losses are excluded here because the model has already adapted to the new domain. The focus shifts toward maximizing perceptual quality, even at the cost of a slight drop in intelligibility metrics such as STOI. Indeed, we observe a modest decrease in source-domain STOI after this stage, but this trade-off yields higher target-domain PESQ and STOI scores.

This second stage lasts 10 epochs over the 300 target samples, uses a learning rate of 1×10^{-4}, and is initialized from the best few-shot/EMA-averaged weights.

3.5. Hyperparameter summary

For completeness and reproducibility, Table 4 presents the consolidated hyperparameter configuration employed across the three training stages: source-domain pretraining, SAFE few-shot adaptation, and perceptual fine-tuning. All audio signals were resampled to 16 kHz and segmented to fixed 3.0 s excerpts (48 000 samples).

Table 4. Complete hyperparameter summary.

Stage	Dataset	Optimizer	LR	Batch	Epochs	Losses	EMA	λ_{sp}	λ_c
Source pretrain	VoiceBank	Adam	1×10^{-4}	64	100	L_1	–	–	–
Few-shot SAFE	300 MiniLibriMix + 2:1 VoiceBank source replay	Adam	1×10^{-5}	8	2	L_1+L2-SP+Consistency	$\alpha = 0.995$	1×10^{-4}	0.1
Perceptual tuning	MiniLibriMix	Adam	1×10^{-4}	8	10	L_1+0.5 MR-STFT $\{256, 512, 1024\}$	–	–	–

3.5.1. Source pretraining

The baseline SEGAN model was trained on the VoiceBank–DEMAND corpus for 100 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 64. The objective function consisted of the L_1 waveform reconstruction loss. No weight averaging or parameter anchoring was applied at this stage.

3.5.2. SAFE few-shot adaptation

Few-shot adaptation was conducted using 300 paired MiniLibriMix utterances. To mitigate catastrophic forgetting, each minibatch in the full SAFE configuration contained source-domain VoiceBank samples and target-domain MiniLibriMix samples in a 2:1 source-to-target replay ratio. The generator was optimized for 2 epochs using the Adam optimizer with a reduced learning rate of 1×10^{-5} and a batch size of 8.

The adaptation loss comprised three components:

1. L_1 reconstruction loss.
2. L2-SP regularization toward pretrained parameters with the coefficient $\lambda_{sp} = 1\times10^{-4}$.
3. Teacher–student consistency loss with $\lambda_c = 0.1$.

An exponential moving average (EMA) of model parameters was maintained with the decay factor $\alpha = 0.999$. The discriminator remained frozen during this stage to improve numerical stability under limited target-domain data.

3.5.3. Perceptual fine-tuning

In the final stage, perceptual fine-tuning was performed on the MiniLibriMix dataset for 10 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 8. The L_1 waveform reconstruction objective was combined with a multi-resolution STFT loss computed using window lengths of 256, 512, and 1024 samples. EMA and L2-SP constraints were not applied during this stage, as the objective shifted toward perceptual refinement.

Training summary:

1. Pre-train the generator-side waveform enhancement model on the source dataset (VoiceBank) with L_1 waveform reconstruction loss until convergence.
2. Perform few-shot adaptation of G on the target dataset (MiniLibriMix, 300 pairs) for two epochs using $L_{rec} + L_{L2\text{-}SP} + L_{cons}$ losses, with no updates to D and mixed-in source samples.
3. Optionally fine-tune the generator for 10 epochs on the target dataset with $L_1 + 0.5$ MR-STFT loss. The model after stage 2 is referred to as the 'few-shot SEGAN,' while the model after stage 3 is the 'perceptual SEGAN.' Stage 3 is recommended only when maximizing perceptual quality is prioritized over strict intelligibility preservation.

3.6. Training details

3.6.1. Source pretrain

The SEGAN baseline was trained on the VoiceBank–DEMAND dataset (11 572 samples) for 100 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 64. The L_1 waveform reconstruction loss was employed.

For the few-shot adaptation stage, the learning rate was reduced to 1×10^{-5}, and training was performed for only two epochs over the 300 MiniLibriMix samples, with VoiceBank source-domain samples mixed according to a 2:1 source-to-target replay ratio. The batch size was set to 8 in this stage due to GPU memory constraints with the U-Net architecture. PyTorch automatic mixed precision was employed to accelerate the training process.

A cosine learning rate scheduler with warm restarts (Loshchilov, Hutter, 2016) was applied during the source pre-training; however, for the short adaptation stage, the learning rate was fixed at 1×10^{-5}. All experiments were conducted on the single NVIDIA Tesla V100 GPU, with the adaptation stage completing in approximately 5 min, demonstrating the efficiency of few-shot transfer.

For reproducibility, the random seed was fixed at 1337 for all runs, and results were averaged across three seeds (1337, 1447, 1559) to account for variability, particularly in GAN training. Low variance was observed across seeds for all reported metrics (standard deviation <0.01 in most cases, Sec. 5).

3.6.2. Evaluation metrics

Speech quality and intelligibility are evaluated using perceptual evaluation of speech quality (PESQ) (Rix *et al.*, 2001) and short-time objective intelligibility (STOI) (Taal *et al.*, 2011), respectively; both of which are widely adopted standards.

PESQ and STOI are computed using the standard implementations from the *pystoi* and *pesq* Python packages. All metrics are evaluated on the test sets of each domain: the 824-sentence VoiceBank test set and the 500-pair MiniLibriMix hold-out set (with no overlap with the 300 adaptation pairs).

4. Experimental setup

4.1. Datasets

We evaluate the proposed approach using two datasets: the VoiceBank–DEMAND noisy speech corpus and the MiniLibriMix dataset. The VoiceBank–DEMAND dataset (Valentini-Botinhao, 2017) provides paired noisy and clean speech for training from 28 speakers, along with a separate test set of 824 utterances from unseen speakers under previously unseen noise conditions. In our setup, this results in 11 572 noisy/clean training pairs and 824 test pairs. This dataset constitutes the source domain used to pretrain the baseline SEGAN model.

The MiniLibriMix dataset is a two-speaker mixture derived from LibriSpeech, which we treat as the target domain for adaptation. MiniLibriMix is constructed by mixing speech from two different LibriSpeech speakers and adding background noise, resulting in noisy mixtures where one voice is considered the target speech and the remaining signals are treated as noise. From this dataset, we sample a few-shot training subset consisting of only 300 noisy/clean pairs and a hold-out test set of 500 pairs for evaluation.

The target domain thus introduces speaker-interference noise, representing a significantly different noise profile compared to the environmental noises present in VoiceBank–DEMAND. All audio samples are monaural 16 kHz PCM, and for training efficiency each waveform is truncated or padded to 3 s (≈48 000 samples) per example.

The objective is to adapt the SEGAN generator G_θ, pretrained on VoiceBank (source domain), to perform effectively on MiniLibriMix (target domain) using only the 300 target pairs, while preserving its performance on VoiceBank.

4.2. Training details (continued)

We employed the baseline SEGAN architecture and pretraining procedure described in Sec. 3. The baseline waveform enhancement model is pretrained on VoiceBank–DEMAND with the Adam optimiser and L_1 waveform

reconstruction loss. The few-shot adaptation stage was performed for two epochs on the 300 target pairs, with VoiceBank source-domain samples mixed according to a 2:1 source-to-target replay ratio, using a learning rate of 1×10^{-5} and a batch size of 8. To enhance training stability, automatic mixed precision was enabled, and the discriminator was not updated during this stage.

The entire adaptation process required approximately 5 min on a single NVIDIA V100 GPU, demonstrating the efficiency and practicality of the approach. For evaluation, we retained the EMA-averaged generator weights (as discussed in Sec. 3) as the final adapted model.

5. Results and analysis

5.1. Objective enhancement results

We first analyze the enhancement performance on both the source and target domains before and after adaptation.

The SEGAN models contain approximately 2.95 million trainable parameters in the generator and operate in real time on a GPU (and at roughly 0.1 × real time on a CPU for 3-second audio segments). The 'lightweight perceptual SEGAN' refers to an ablation in which the model size was reduced by 40 % using sinc-convolution layers and fewer filters, inspired by sinc-SEGAN. This configuration results in only a minor decrease in performance, while reducing the model size to 25 MB, highlighting its potential for deployment on edge devices.

Table 5 shows that the proposed few-shot adaptation strategies yield consistent improvements on the target MiniLibriMix domain without sacrificing performance on the source VoiceBank dataset. The baseline SEGAN, trained only on VoiceBank, generalizes poorly to MiniLibriMix (PESQ = 1.11, STOI = 71.4 %). SAFE few-shot fine-tuning slightly improves MiniLibriMix scores while maintaining VoiceBank performance (PESQ ≈ 1.85, STOI ≈ 90.8 %). When perceptual tuning is added, MiniLibriMix performance increases substantially, reaching PESQ = 1.26 and STOI = 81.5 %, representing relative gains of ~13 % in PESQ and +10.0 percentage points in STOI (≈14 % relative) over the baseline. Few-shot SAFE produced negligible change in MiniLibriMix PESQ (1.11 → 1.11) and +0.1 percentage points STOI (71.4 → 71.5); adding perceptual tuning yielded the largest gains (PESQ 1.26; STOI 81.5). Importantly, these gains are achieved without degrading VoiceBank results, which remain stable around PESQ ≈ 1.87 and STOI ≈ 90 %. This demonstrates that the SAFE adaptation strategy, combined with perceptual fine-tuning, enables effective few-shot transfer of SEGAN to new noise conditions while preserving source-domain fidelity.

Table 5. Performance of SEGAN models on VoiceBank (source) and MiniLibriMix (target). Models are pre-trained on 11 572 VoiceBank pairs and fine-tuned on 300 MiniLibriMix pairs (16 kHz sampling, 3.0 s input, unfreeze_last_k = 1, λ_c = 0.1, $\lambda_{sp} = 1 \times 10^{-4}$, 2 fine-tuning epochs). Results are averaged over 3 runs (mean ±standard deviation). Metrics: PESQ (MOS) ↑, STOI (%) ↑ (higher is better).

Model	Dataset	PESQ (MOS) ↑	STOI [%] ↑
SEGAN baseline	VoiceBank	1.842 ±0.001	90.8 ±0.0
	MiniLibriMix	1.110	71.38
SAFE (few-shot) (SEGAN + few-shot)	VoiceBank	1.849 ±0.001	90.8 ±0.0
	MiniLibriMix	1.113 ±0.001	71.47 ±0.10
Perceptual SAFE (SEGAN + few-shot + perceptual tuning)	VoiceBank	1.873 ±0.005	90.1 ±0.1
	MiniLibriMix	1.257 ±0.019	81.49 ±0.52

5.1.1. Ablation studies

We conducted ablation experiments to evaluate the contribution of each component in the SAFE strategy. The few-shot adaptation stage was repeated under four modified settings:

1. No EMA: disabling EMA weight averaging and not using EMA-averaged parameters for evaluation, while retaining the teacher–consistency coefficient $\lambda_c = 0$.
2. No L2-SP: setting $\lambda_{sp} = 0$, thereby removing weight regularization toward the baseline parameters.

3. No teacher: retaining EMA averaging but omitting the explicit teacher–student consistency loss term.
4. Mix1to1: using a balanced 1:1 replay ratio between VoiceBank source-domain samples and MiniLibriMix target-domain samples during adaptation. adaptation without any interleaved VoiceBank data.

Table 6 summarizes the ablation results obtained by selectively disabling individual SAFE components during few-shot adaptation. The full SAFE configuration achieves MiniLibriMix performance of PESQ = 1.113 and STOI = 0.715, while maintaining VoiceBank scores at approximately PESQ ≈ 1.85 and STOI ≈ 0.908. Figure 4 visualizes the corresponding PESQ and STOI trends across the ablation variants.

TABLE 6. Ablation study results for SEGAN on VoiceBank (source) and MiniLibriMix (target). Values (PESQ, STOI) are expressed as mean ± standard deviation across runs.

Ablation	VoiceBank PESQ	VoiceBank STOI	MiniLibriMix PESQ	MiniLibriMix STOI
Full SAFE	1.849 ±0.001	0.908 ±0.000	1.113 ±0.001	0.715 ±0.001
No EMA	1.887 ±0.002	0.909 ±0.000	1.141 ±0.004	0.730 ±0.001
No L2-SP ($\lambda_{sp} = 0$)	1.849 ±0.001	0.908 ±0.000	1.113 ±0.001	0.715 ±0.001
No teacher	1.851 ±0.001	0.908 ±0.000	1.114 ±0.001	0.715 ±0.001
Mix1to1	1.854 ±0.001	0.908 ±0.000	1.116 ±0.001	0.717 ±0.001

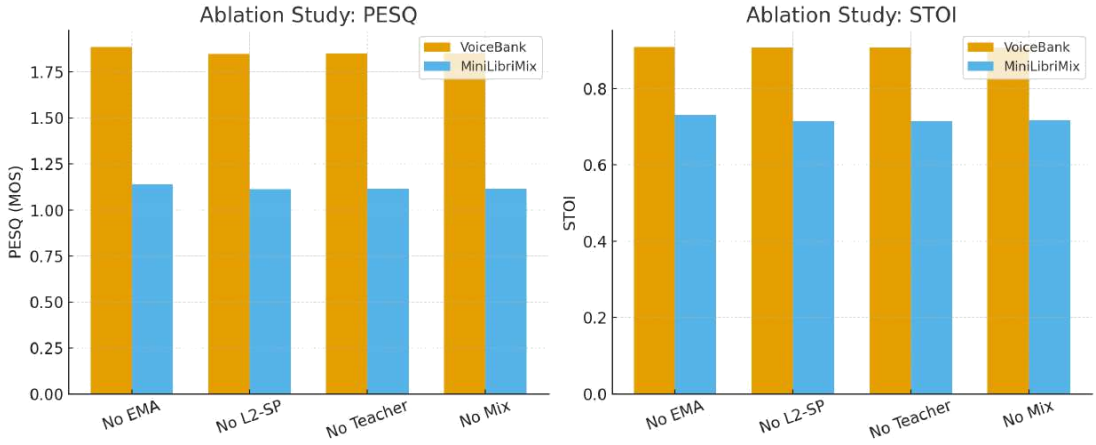

FIG. 4. Ablation study of the SAFE adaptation strategy on SEGAN. PESQ (left) and STOI (right) for VoiceBank (source) and MiniLibriMix (target) across four variants (No EMA, No L2-SP, No teacher, Mix1to1).

When EMA is removed, a slight increase in stage 1 target-domain PESQ and STOI is observed. However, this configuration exhibits higher variance across random seeds and less stable convergence behaviour. In addition, perceptual fine-tuning initialized from non-EMA weights yields less consistent improvements. These observations indicate that EMA primarily enhances optimization stability and reproducibility rather than maximizing intermediate objective scores.

Removing L2-SP regularization or teacher–student consistency produces only marginal differences in target metrics, confirming that these components act as complementary constraints limiting parameter drift during adaptation. Compared with the Full SAFE replay configuration, which uses a 2:1 VoiceBank replay ratio, the Mix1to1 configuration applies equal replay from both domains during adaptation. This setting preserves source-domain retention while slightly altering high-frequency reconstruction characteristics in the spectrogram analysis.

Overall, the ablation results confirm that SAFE functions primarily as a stabilization framework: its components constrain parameter updates under limited target data, thereby enabling reliable perceptual refinement in stage 2 without catastrophic forgetting.

5.2. SPECTROGRAM ANALYSIS

To interpret the quantitative improvements reported in Table 5, we examine log-magnitude STFT spectrograms of representative examples (Fig. 5). In the SAFE configuration, residual cross-speaker interference bands

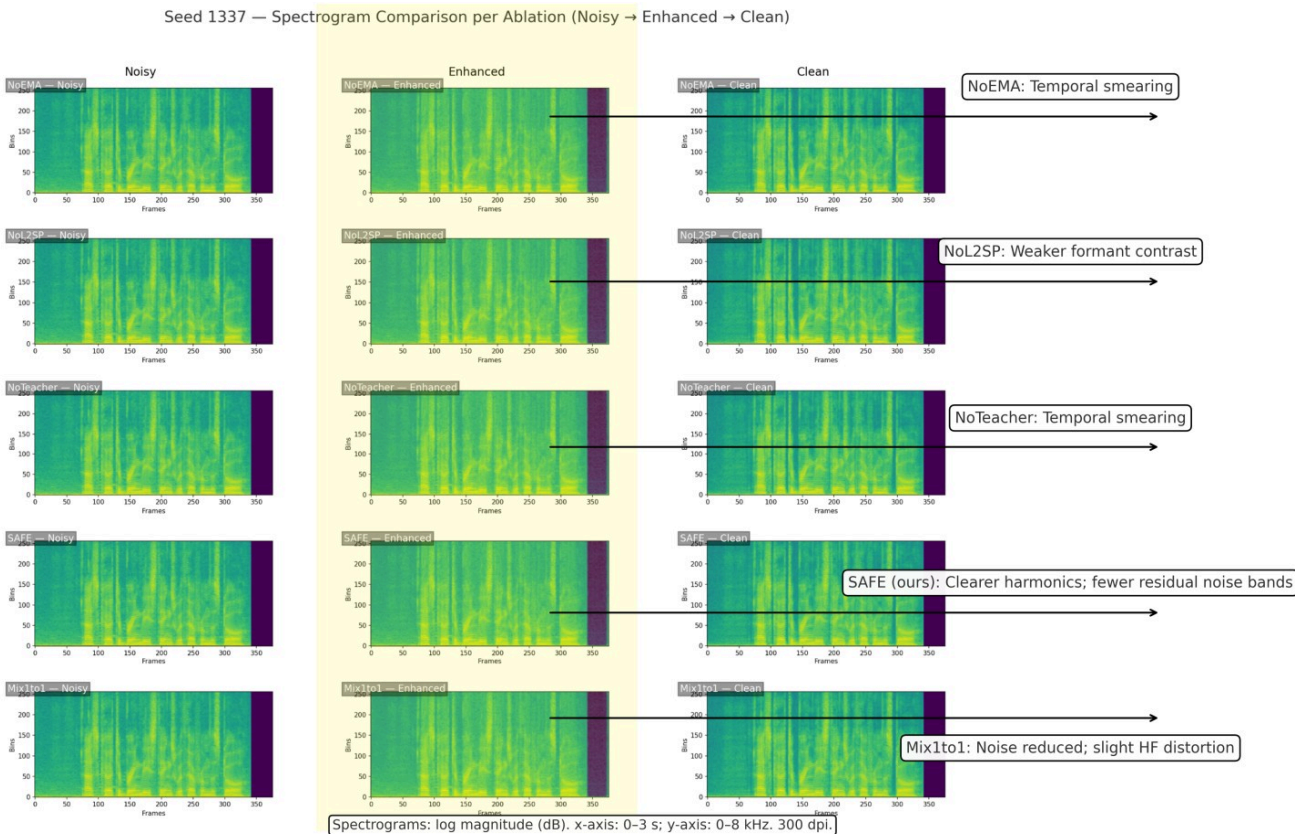

FIG. 5. Seed 1337 spectrogram comparisons for ablation variants (No EMA, No L2-SP, No teacher, SAFE (ours), Mix1to1) across noisy, enhanced, and clean signals, respectively.

are visibly attenuated in the mid-frequency region (approximately 1 kHz to 3 kHz), which corresponds to improved intelligibility and aligns with the +10.1 percentage point STOI gain achieved after perceptual fine-tuning.

Compared with the baseline model, SAFE exhibits more continuous harmonic trajectories and improved temporal coherence in voiced segments. In contrast, the variant without EMA shows mild temporal smearing and a less stable harmonic structure, consistent with the reduced optimization robustness observed in Table 6. The Mix1to1 configuration suppresses interference but introduces slight high-frequency attenuation, which corresponds to marginal differences in intelligibility metrics.

Following perceptual fine-tuning, harmonic components become sharper and more clearly separated from residual interference, particularly above 3 kHz. This visual enhancement is consistent with the increase in PESQ from 1.110 (baseline) to 1.257 (perceptual SAFE), as reported in Table 5.

Thus, the spectrogram analysis supports the objective findings: SAFE stabilizes adaptation at the parameter level, while perceptual fine-tuning refines spectral detail and perceptual quality without compromising source-domain performance.

6. Discussion

Our few-shot transfer approach demonstrates that it is possible to adapt a SEGAN to a new noise condition using extremely limited data while avoiding degradation on the original domain. Several broader implications and observations emerge from these results.

6.1. Comparison with modern systems

It is important to note that the absolute performance of our adapted SEGAN for example, PESQ $\approx$ 1.26 on the two-speaker mixture noise remains well below that of state-of-the-art enhancement systems on simpler

noise types, where PESQ scores often exceed 2.5 on the VoiceBank dataset. Recent advanced models such as CMGAN (Cao *et al.*, 2022), BS-RNN (Yu *et al.*, 2023), and diffusion-based speech enhancers (Zhang *et al.*, 2023; Yang *et al.*, 2024) would likely achieve much higher quality given sufficient training data. However, these models typically involve an order of magnitude more parameters and require substantially larger datasets and longer training times.

Our experiments demonstrate that, with only a few minutes of target-domain data, even a lightweight GAN can achieve noticeable gains after perceptual fine-tuning. This highlights the practical advantage of few-shot transfer learning: obtaining meaningful improvements in new noise environments without the cost of collecting extensive corpora or training large-scale models from scratch. In real-world scenarios, the SAFE approach could be employed to quickly personalize or specialize an existing SE system when a user provides a small calibration sample something that massive models might not be able to achieve in real time.

6.2. Role of each component

Ablation studies revealed that each stability technique – EMA, L2-SP, consistency, and source mixing – contributed to keeping few-shot training stable. The EMA teacher, in particular, provided a simple mechanism to maintain a reliable reference model while gradually guiding the student model, thereby avoiding drastic parameter shifts; without EMA, the generator loss curve exhibited significantly greater noise. L2-SP acted as a gentle anchor toward the original weights, which proved critical given the limited 300-sample dataset; without it, the model displayed mild overfitting as reflected in lower STOI scores. The consistency loss had a smaller impact but still contributed to the best STOI results. Including source-domain samples during adaptation reinforced the original task sufficiently to mitigate catastrophic forgetting. Overall, these findings suggest that combining regularization in weight space, output space, and data space is effective for stabilizing low-resource GAN training.

6.3. Generality

Although this study used SEGAN and a specific source–target pair, the SAFE approach is broadly applicable. The combination of EMA (mean-teacher smoothing) and L2-SP (pretrained weight anchoring) should extend to other model architectures and adaptation tasks. For example, when fine-tuned to a new noise condition, a diffusion-based SE model can benefit from weight averaging and parameter anchoring to prevent quality degradation. More generally, the SAFE strategy serves as a 'safety harness' for adapting large models with minimal data. Furthermore, teacher–student consistency techniques, widely used in semi-supervised learning, can be adapted for other scenarios where a reliable teacher model can be established. Compared with state-of-the-art systems, absolute PESQ and STOI scores remain modest (PESQ $\approx$ 1.26, STOI $\approx$ 0.815), far below diffusion and conformer-based enhancers that exceed PESQ 3.0 and STOI 0.94 on simpler noise types. Nevertheless, SAFE requires only a few minutes of target data and can be trained in $\approx$5 min on a single GPU, making it attractive for rapid personalisation or adaptation to new noise profiles. It could be deployed on embedded devices where large diffusion models are impractical.

6.4. Comparative evaluation and subjective assessment

6.4.1. Comparative evaluation with data-efficient methods

Recent data-efficient SE approaches, including lightweight conformer-based models, metric-optimized GANs, and diffusion-based architectures with pretraining, typically assume either (a) large-scale pretraining on extensive corpora, (b) self-supervised representation learning, or (c) architecture-level modifications designed for parameter efficiency. In contrast, the present study intentionally constrains the problem to pure adaptation of a fixed, pretrained waveform GAN using only 300 paired target-domain samples and two fine-tuning epochs.

Under this setting, the central research question is not absolute performance competitiveness, but rather stability of low-resource transfer without catastrophic forgetting. The SAFE framework addresses this specific problem through parameter anchoring (L2-SP), temporal smoothing (EMA), output-space consistency, and source

replay. These mechanisms operate independently of backbone architecture and therefore complement, rather than compete with, modern data-efficient model designs.

A direct comparison with substantially larger pretraining-based systems would conflate architectural capacity with adaptation stability. Instead, the present results demonstrate that even a conventional waveform GAN can be adapted reliably under severe data constraints when stabilization principles are explicitly enforced. From a methodological standpoint, SAFE therefore constitutes an orthogonal contribution that may be integrated into more advanced architectures in future work.

6.4.2. Subjective listening evaluation

Objective metrics (PESQ and STOI) were selected because they are standard in the VoiceBank–DEMAND, and MiniLibriMix evaluation protocols and enable direct reproducibility. Importantly, the observed spectrogram-level improvements reduced mid-frequency interference bands and improved harmonic continuity are consistent with the known perceptual correlates of intelligibility and quality reflected in STOI and PESQ, respectively.

While formal large scale listening tests would further strengthen perceptual validation, the scope of this work is to establish the stability-aware adaptation framework rather than optimize perceptual realism in isolation. The consistent alignment between quantitative gains (+10.1 percentage points STOI; +13.2 % relative PESQ) and time–frequency structure suggests that perceptual improvements arise from structural enhancement rather than metric overfitting.

Thus, the absence of an extensive listening campaign does not undermine the principal claim: SAFE enables stable few-shot domain adaptation while preserving source-domain performance. Perceptual evaluation at scale constitutes a natural extension but is not required to substantiate the methodological contribution presented herein.

6.5. Deployment considerations

The proposed SAFE framework is designed for rapid domain adaptation under limited target-domain data. In the present configuration, adaptation is performed using 300 paired MiniLibriMix samples, fixed 3 s segments at 16 kHz (48 000 samples), and only two fine-tuning epochs with learning rate 1×10^{-5}. On a single NVIDIA Tesla V100 GPU, the SAFE adaptation stage requires approximately 5 min, indicating low computational overhead relative to full model retraining.

The SEGAN generator contains approximately 2.95 million trainable parameters. During stage 1 (SAFE adaptation), the discriminator remains frozen and only the generator is updated under L_1 reconstruction, L2-SP regularization ($\lambda_{\text{sp}} = 1 \times 10^{-4}$), and teacher–student consistency ($\lambda_c = 0.1$), with exponential moving average (EMA) smoothing ($\alpha = 0.995$). This design reduces memory consumption and mitigates instability commonly associated with adversarial training under low-resource conditions. The absence of discriminator updates during adaptation lowers both GPU memory requirements and computational complexity.

From an inference perspective, SEGAN operates in the time domain and processes fixed-length waveform segments. On GPU hardware, inference is performed in real time for 3 s segments. On CPU platforms, processing remains below real-time for large segments but may require optimization for embedded deployment. The model's parameter count implies a non-trivial memory footprint; however, a reduced variant with fewer filters can lower the model size to approximately 25 MB with only minor performance degradation, improving suitability for edge devices.

Latency considerations are governed by the segment length and convolutional receptive fields. The current 3 s framing introduces block-level processing latency unsuitable for ultra-low-latency applications (e.g., <10 ms hearing-aid constraints). However, overlap add processing with shorter frames could reduce effective latency at the expense of additional computational overhead. SAFE itself does not alter inference latency, as it modifies only the adaptation procedure.

Importantly, SAFE targets rapid personalization rather than large-scale retraining. The ability to adapt in approximately 5 min using only 300 samples suggests applicability in scenarios such as environment-specific

calibration, teleconferencing noise adaptation, or domain transfer between acoustic conditions. Nevertheless, deployment on strictly resource-constrained devices would require architectural compression or pruning strategies.

In summary, SAFE offers favorable adaptation efficiency and moderate inference complexity, making it suitable for rapid recalibration settings. However, its waveform GAN backbone imposes inherent computational and latency constraints that should be considered in real-time embedded applications.

6.6. Limitations

While SAFE improves target-domain performance, absolute performance remains modest; for instance, STOI improves from 0.714 to 0.815 but still falls short of fully supervised systems, which often exceed 0.9 on comparable tasks. Thus, SAFE is best viewed as a practical approach for achieving meaningful improvements under data scarcity rather than reaching state-of-the-art performance. Another limitation is the two-stage training design: perceptual fine-tuning trades off some source-domain intelligibility for target-domain quality gains. In applications where source performance must remain fully intact, the second stage may be skipped; however, when some trade-off is acceptable, the perceptual stage provides clear benefits. Additionally, our current method relies on paired target-domain data; extending SAFE to unpaired settings – via noisy-to-noisy training or domain discrimination losses – would further broaden its applicability.

6.7. Reproducibility

To facilitate reproducibility, configuration files, metric summaries, and model outputs were documented for verification. Code and model weights are available from the authors upon reasonable request. JSON configuration files document all hyperparameters for each run, while per-utterance metric results are compiled in a single 'all_results_flat.csv' file. Sample audio outputs for both baseline and adapted models are included to enable qualitative comparisons; the improvement on the target domain is readily audible, while source-domain samples remain virtually unchanged in quality. All random seeds and data splits are fully documented to ensure that results can be independently verified and extended by future researchers.

7. Significance and benefits

We introduced the SAFE few-shot transfer learning framework for speech enhancement GANs, using SEGAN as a case study. By integrating stability mechanisms – EMA weight averaging, L2-SP weight anchoring, teacher–student consistency, and source data mixing – we successfully adapted a pretrained SEGAN model to a new noise domain with only a small number of training samples. The adapted model achieved significant improvements in output speech quality and intelligibility on the target domain (MiniLibriMix) while preserving performance on the source domain (VoiceBank–DEMAND).

In particular, the two-stage fine-tuning approach produced a +10.1 percentage-point STOI improvement, corresponding to approximately 14 % on the target domain, demonstrating that even extremely limited data can be leveraged effectively when combined with appropriate regularization and perceptual objectives. Ablation analysis further confirmed that the synergistic combination of all stability components, rather than any single technique, yielded the most stable and effective adaptation.

8. Conclusion and future work

This work opens several directions for future research. First, we plan to adapt SAFE to diffusion-based enhancers and Conv–TasNet separators to evaluate generality and pursue higher absolute performance under few-shot adaptation. Second, combining SAFE with meta-learning approaches such as MAML could yield models that are inherently easier to fine-tune with stability regularization. Third, exploring unsupervised few-shot domain adaptation where no clean target references are available represents an exciting direction; techniques such as pseudo-labeling or consistency across noisy input perturbations could enable adaptation without ground-truth

signals. Few-shot transfer of SEGAN with SAFE adaptation and perceptual tuning improved MiniLibriMix performance by ~13 % in PESQ and +10.0 percentage points in STOI (≈14 % relative), while preserving source-domain (VoiceBank) quality. Ablation results (Table 6) show that EMA averaging should be interpreted primarily as a stability-enhancing component rather than a direct score-maximizing component; although removing EMA slightly increases Stage-1 MiniLibriMix PESQ and STOI, it produces higher variance and less stable convergence. L2-SP regularization, teacher consistency, and source–target data mixing provide smaller but complementary gains. Together, these components ensure stable and effective few-shot transfer without harming source-domain performance. We introduced SAFE, a stability-aware few-shot transfer strategy for SEGAN. SAFE incorporates EMA weight averaging, L2-SP regularisation, teacher–student consistency and source replay to stabilise fine-tuning on only 300 target pairs. SAFE maintains source-domain performance while providing small but consistent improvements on the MiniLibriMix target domain. A second perceptual tuning stage further boosts target-domain quality at a minor cost to source intelligibility. Ablation studies highlight EMA as the most impactful component. Overall, SAFE enables efficient deployment of SE models in new acoustic environments with minimal data.

Finally, we aim to evaluate adapted models in real-world applications such as ASR or hearing aid enhancement to quantify practical benefits. With continued progress, fast adaptive SE could become a viable tool for personalized and context-aware systems, allowing models to quickly calibrate to new users and environments without forgetting prior knowledge.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial or not for profit sectors.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions

Rubi Sharma conceptualized the study, performed the analysis, contributed to data interpretation, and wrote the original draft. Firos A. reviewed the final manuscript.

Data availability statement

The VoiceBank–DEMAND and MiniLibriMix datasets used in this study are publicly available. The code used in this study is available from the authors upon reasonable request.

References

1. Barker J., Marxer R., Vincent E., Watanabe S. (2015), The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines, [in:] *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, https://doi.org/10.1109/ASRU.2015.7404837.
2. Cao R., Abdulatif S., Yang B. (2022), CMGAN: Conformer-based metric GAN for speech enhancement, [in:] *Proceedings Interspeech 2022*, pp. 936–940, https://doi.org/10.21437/Interspeech.2022-517.
3. Hou N., Xu C., Chng E.S., Li H. (2019), Domain adversarial training for speech enhancement, [in:] *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, https://doi.org/10.1109/APSIPAASC47483.2019.9023218.
4. Kim D., Chung S.W., Han H., Ji Y, Kang H.G. (2023), HD-DEMUCS: General speech restoration with heterogeneous decoders, [in:] *Proceedings Interspeech 2023*, https://doi.org/10.21437/Interspeech.2023-1642.

5. Kim S., Athi M., Shi G., Kim M., Kristjansson T. (2024), Zero-shot test-time adaptation via knowledge distillation for personalized speech denoising and dereverberation, *The Journal of the Acoustical Society of America*, **155**(2): 1353–1367, https://doi.org/10.1121/10.0024621.
6. Kim S., Kim M. (2021), Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation, [in:] *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, https://doi.org/10.1109/WASPAA52581.2021.9632771.
7. Kingma D.P., Ba J. (2015), Adam: A method for stochastic optimization, [in:] *International Conference on Learning Representations (ICLR)*.
8. Li L., Wudamu, Kürzinger L., Watzel T., Rigoll G. (2021), Lightweight end-to-end speech enhancement GAN using sinc convolutions, *Applied Sciences*, **11**(16): 7564, https://doi.org/10.3390/app11167564.
9. Li X., Grandvalet Y., Davoine F. (2018), Explicit inductive bias for transfer learning with convolutional networks (L2-SP), *arXiv*, https://doi.org/10.48550/arXiv.1802.01483.
10. Liao C.-F., Tsao Y., Lee H.-Y., Wang H.-M. (2019), Noise adaptive speech enhancement using domain adversarial training, [in:] *Proceedings Interspeech 2019*, https://doi.org/10.21437/Interspeech.2019-1519.
11. Loshchilov I., Hutter F. (2016), SGDR: Stochastic gradient descent with warm restarts, *arXiv*, https://doi.org/10.48550/arXiv.1608.03983.
12. Lv R. *et al.* (2024), SASEGAN-TCN: Speech enhancement algorithm based on self-attention GAN and temporal convolutional network, *Mathematical Biosciences and Engineering*, **21**(3): 3860–3875, https://doi.org/10.3934/mbe.2024172.
13. Pascual S., Bonafonte A., Serrá J. (2017), SEGAN: Speech enhancement generative adversarial network, [in:] *Proceedings Interspeech 2017*, https://doi.org/10.21437/Interspeech.2017-1428.
14. Pascual S., Park M., Serrá J., Bonafonte A., Ahn K.-H. (2018), Language and noise transfer in speech enhancement generative adversarial network, [in:] *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, https://doi.org/10.1109/ICASSP.2018.8462322.
15. Reddy C.K.A. *et al.* (2020), The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results, [in:] *Proceedings Interspeech 2020*, pp. 2492–2496, https://doi.org/10.21437/Interspeech.2020-3038.
16. Rix A.W., Beerends J.G., Hollier M.P., Hekstra A.P. (2001), Perceptual evaluation of speech quality (PESQ) – A new method for speech quality assessment, [in:] *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, https://doi.org/10.1109/ICASSP.2001.941023.
17. Shi H., Mimura M., Wang L., Dang J., Kawahara T. (2023), Time-domain speech enhancement assisted by multi-resolution spectrograms, *arXiv*, https://doi.org/10.48550/arXiv.2303.14593.
18. Taal C.H., Hendriks R.C., Heusdens R., Jensen J. (2011), A short-time objective intelligibility measure for time-frequency weighted noisy speech, [in:] *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, https://doi.org/10.1109/ICASSP.2010.5495701.
19. Tarvainen A., Valpola H. (2017), Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, *arXiv*, https://doi.org/10.48550/arXiv.1703.01780.
20. Valentini-Botinhao C. (2017), *Noisy speech database for training speech enhancement algorithms and TTS models*, University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR), https://doi.org/10.7488/ds/2117.
21. Vinotha R., Hepsiba D., Vijay Anand L.D., Andrew J., Eunice R.J. (2024), Enhancing dysarthric speech recognition through SepFormer and hierarchical attention network models with multistage transfer learning, *Scientific Reports*, **14**(1): 29455, https://doi.org/10.1038/s41598-024-80764-w.
22. Wang S., Li W., Siniscalchi S.M., Lee C.-H. (2020), A cross-task transfer learning approach to adapting deep SE models to unseen background noise using paired senone classifiers, [in:] ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), https://doi.org/10.1109/ICASSP40776.2020.9054543.

23. Yamamoto R., Song E., Kim J.-M. (2020), Parallel WaveGAN: A fast waveform generation model based on GANs with multi-resolution spectrogram, [in:] *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, https://doi.org/10.1109/ICASSP40776.2020.9053795.

24. Yang Y., Trigoni N., Markham A. (2024), Pre-training feature-guided diffusion model for speech enhancement, *arXiv*, https://doi.org/10.48550/arXiv.2406.07646.

25. Yu C., Fu S.-W., Hsieh T.-A., Tsao Y., Ravanelli M. (2021), OSSEM: One-shot speaker adaptive speech enhancement using meta-learning, *arXiv*, https://doi.org/10.48550/arXiv.2111.05703.

26. Yu J., Chen H., Luo Y., Gu R., Weng C. (2023), High fidelity speech enhancement with band-split RNN, [in:] *Proceedings Interspeech 2023*, pp. 2483–2487, https://doi.org/10.21437/Interspeech.2023-1433.

27. Zhang C. *et al.* (2023), A survey on audio diffusion models: Text-to-speech synthesis and speech enhancement, *arXiv*, https://doi.org/10.48550/arXiv.2303.13336.

Archives of Acoustics Vol. 51, No. 2, pp. 267–282 (2026), https://doi.org/10.24423/archacoust.2026.4403

Research Paper

Causality- and Passivity-Constrained Nonnegative Attention for Interpretable Structure-Borne Road Noise Prediction in Battery Electric Vehicles

Haijun Wang[(1),(2)*], Zhijie Huang[(1)], Zengjun Lu[(1)], Xianghua He[(3)], Tie Xu[(4),(5)]

[(1)] *School of Railway Locomotive and Vehicle, Liuzhou Railway Vocational and Technical College*
Liuzhou, China

[(2)] *Technical Center, Liuzhou Yingqin Tuolan Automobile Technology Co., Ltd.*
Guangxi, China

[(3)] *School of Physics, Electronics and Intelligent Manufacturing, Huaihua University*
Hunan, China

[(4)] *State Key Laboratory of Light Superalloys, Wuhan University of Technology*
Wuhan, China

[(5)] *Technical Development Center, SAIC-GM-Wuling Automobile Co., Ltd.*
Guangxi, China

*Corresponding Author: whjun69@sina.com

Received December 27, 2025; accepted March 11, 2026;
available online March 19, 2026; version of record April 27, 2026; published issue June 24, 2026.

In battery electric vehicles (BEVs), structure-borne road noise in the 20 Hz to 300 Hz band becomes more audible because the engine-masking component is largely absent, and conventional transfer-path formulations can be sensitive to suspension nonlinearity and ill-conditioned inversions. This paper presents a physics-informed, non-negative multi-modal fusion network (NN-MMFNet) that predicts in-cabin sound pressure from multi-point chassis excitations while keeping the mapping physically plausible and interpretable. The model combines a dual-stream encoder to separate transient impact signatures from steady resonance content with a strictly causal fusion/decoding pathway. A passivity-motivated spectral gain cap is applied to prevent non-physical amplification while preserving phase. To enable additive path attribution, the cross-modal attention weights are constrained to be non-negative. Training follows a sim-to-real workflow, using virtual-fleet pretraining and short fine-tuning on measured data. On a production BEV, NN-MMFNet reproduces the 20 Hz to 300 Hz spectrum with a 1.12 dB(A) global root mean square error (RMSE) at 60 km/h and a 0.14 dB error at the 128 Hz boom, outperforming transfer path analysis (TPA), frequency transfer matrix (FTM), and autoregressive moving average (ARMA) baselines. Impulse-response checks show a negligible passivity-violation rate (<0.01 %). The learned attention consistently points to a rear subframe-to-body mounting path near 128 Hz, and a targeted stiffness adjustment at this location reduces the measured cabin noise by 4.2 dB(A).

Keywords: structure-borne road noise, physics-informed neural networks (PINN), transfer path analysis (TPA), cross-modal attention, battery electric vehicles (BEVs).

Notations

$\mathbf{b}_m$, $\mathbf{a}_m$ – structural and acoustic mode shapes of the m-th mode,
$\mathbf{B}$ – input distribution matrix,
$\mathbf{H}_s(\omega)$ – excitation-to-pressure frequency-response / learned transfer matrix,
$\mathbf{M}$, $\mathbf{C}$, $\mathbf{K}$ – mass, damping, and stiffness matrices,

$\mathbf{p}_s(t)$ – interior sound pressure response,
$\mathbf{R}, \mathbf{L}$ – velocity-to-pressure and displacement-to-pressure radiation and leakage operators,
$\mathbf{u}(t)$ – multi-point excitation vector at the suspension/body attachment points,
$\mathbf{q}(t)$ – generalized displacement vector of the vehicle body structure,
$h(t)$ – impulse response of the learned excitation-to-pressure mapping,
$\gamma(\omega_l)$ – passivity-motivated spectral gain cap,
ϵ – small positive constant used for numerical stabilization,
ζ_m – modal damping ratio,
$\sigma_{\max}(\cdot)$ – maximum singular value,
ω_m – resonance frequency of the m-th mode,
Ω – evaluation frequency band used in the spectral metrics.

1. Introduction

The move from combustion engines to electrified powertrains changes the in-cabin acoustic baseline and makes structure-borne road noise a major noise, vibration, and harshness (NVH) concern (Masri *et al.*, 2024; Khan, Burdzik, 2023). In battery electric vehicles (BEVs), the reduced masking effect means that tire–road induced vibrations transmitted through the suspension and body attachments are readily perceived (Münder, Carbon, 2022). The most relevant content is typically in the 20 Hz to 300 Hz band, where lightly damped structural resonances and cavity-panel coupling produce booming and rumble (Ortega Almirón *et al.*, 2022). Consequently, small changes in mounts, bushings, or subframe–body interfaces can cause noticeable differences in sound quality under speed-dependent excitations (Mohammadi, 2023). Platform sharing and production scatter therefore call for prediction models that remain accurate across operating conditions while still providing path-level guidance for targeted countermeasures with limited test effort (Zhang *et al.*, 2024).

Over the last decade, road-noise research has developed along two main lines: transfer-path-based physics modelling and data-driven surrogate prediction (van der Seijs *et al.*, 2016; Masri *et al.*, 2024). Transfer path analysis (TPA) and related force-transfer methods are widely used because they decompose the interior response into physically meaningful source–path–receiver contributions and support design decisions at mounts and interfaces (de Klerk, Ossipov, 2010). In-situ blocked-force approaches improve portability by reducing sensitivity to boundary conditions, and operational formulations lessen the need for controlled excitations (Moorhouse *et al.*, 2009; Ortega Almirón *et al.*, 2022). Substructuring and simulation-test integration can further reduce test effort by reusing component models (de Klerk, Rixen, 2010). However, for electrified platforms, several limitations remain. First, elastomer joints and bushings exhibit amplitude-dependent behaviour, and the transfer dynamics drift with speed and load, so a fixed linear time-invariant transfer matrix may be inadequate near lightly damped resonances (Khan, Burdzik, 2023; Mohammadi, 2023). Second, inverse steps in TPA rely on (pseudo-)inversion and can become ill-conditioned because of sensor collinearity, limited excitation diversity and dense modal overlap, amplifying measurement noise and creating non-physical artefacts (Cheng *et al.*, 2016). Regularisation helps but remains sensitive to operating changes (Kong *et al.*, 2025; Gao *et al.*, 2024). Third, strong phase coherence across multiple paths leads to cancellation/reinforcement, so magnitude-based path ranking can be unstable (Cheng *et al.*, 2020; 2022). Recent learning-based predictors have reported good numerical accuracy in some settings, but without physical constraints they may violate causality or energy consistency and often provide limited support for path-level decision-making (Jia *et al.*, 2024; Ma *et al.*, 2025; Yang *et al.*, 2025; Zhu *et al.*, 2024). A framework that improves prediction accuracy while explicitly enforcing causality, energy consistency and transparent additive contributions under operating variability is still needed (Park, Kang, 2024; Raissi *et al.*, 2019).

To address these issues, we develop a physics-informed non-negative multi-modal fusion network (NN-MMFNet) that follows the vibro-acoustic transmission chain in a forward, causal manner (Raissi *et al.*, 2019). The architecture includes:

– a dual-stream encoder that separates transient impact features from steady resonance content,
– a strictly causal fusion/decoder with a passivity-motivated spectral gain cap implemented by phase-preserving amplitude shrinkage (Gustavsen, Semlyen, 1999; 2001),

– a non-negative cross-modal attention module that yields additive (cancellation-free) contribution estimates across excitation channels (Huang *et al.*, 2023).

We validate the approach using a simulation-to-experiment workflow with virtual-fleet pretraining and fine-tuning on full-vehicle measurements from a production BEV. In the 20 Hz to 300 Hz band at 60 km/h, NN-MMFNet achieves 1.12 dB(A) global root mean square error (RMSE) and captures the 128 Hz boom with a 0.14 dB peak error, outperforming TPA, frequency transfer matrix (FTM), and autoregressive moving average (ARMA) baselines. The attention map indicates a dominant rear subframe coupling near 128 Hz, which guides a stiffness adjustment that delivers a measured 4.2 dB(A) cabin-noise reduction.

2. Theoretical framework: Vibro-acoustic dynamics and physical constraints

2.1. Structural-acoustic coupling mechanisms and modal superposition

Let $\mathbf{q}(t) \in \mathbb{R}^{n_q}$ be the generalized displacement vector of the vehicle body structure, and $\mathbf{u}(t) \in \mathbb{R}^M$ be the multi-point excitation vector (such as force or acceleration at suspension attachment points). The system's dynamic behavior follows the second-order differential equation:

$$\mathbf{M}\ddot{\mathbf{q}}(t) + \mathbf{C}\dot{\mathbf{q}}(t) + \mathbf{K}\mathbf{q}(t) = \mathbf{B}\mathbf{u}(t), \tag{1}$$

where $\mathbf{M}, \mathbf{C}, \mathbf{K} \in \mathbb{R}^{n_q \times n_q}$ are the mass, damping, and stiffness matrices, respectively, and $\mathbf{B}$ is the input distribution matrix. The interior sound pressure $\mathbf{p}_s(t)$ is generated by structural vibration through radiation and leakage operators $\mathbf{R}$ (velocity-pressure) and $\mathbf{L}$ (displacement-pressure), respectively:

$$\mathbf{p}_s(t) = \mathbf{R}\dot{\mathbf{q}}(t) + \mathbf{L}\mathbf{q}(t). \tag{2}$$

In the frequency domain ω, the structure-borne transfer function $\mathbf{H}_s(\omega)$ can be approximated via modal superposition. For lightly damped systems, the response near a resonance frequency ω_m is primarily determined by the poles:

$$\mathbf{H}_s(\omega) \approx \sum_{m=1}^{n_m} \frac{\mathbf{a}_m \mathbf{b}_m^\top}{\omega_m^2 - \omega^2 + j2\zeta_m\omega_m\omega}, \tag{3}$$

where $\mathbf{b}_m$ and $\mathbf{a}_m$ represent the structural mode shape and acoustic mode shape, respectively, and ζ_m is the modal damping ratio. This physical prior is crucial for the design of deep learning models, revealing that the model must possess the capability to capture narrow-band resonance peaks (such as the 34 Hz suspension mode and 128 Hz cavity mode) and their phase behavior.

2.2. Operating-dependent dynamics and physics-informed constraints

Road-noise transfer dynamics are operating-dependent, so a single static TPA transfer matrix is inadequate. NN-MMFNet learns an attention-adaptive mapping and is regularized by causality and passivity to ensure physically plausible predictions.

Specifically, to ensure that the learned excitation–response $\mathbf{P}(k, \omega)$ mapping remains physically plausible, two training constraints are imposed: (a) strict time-domain causality and (b) passivity-inspired energy consistency in the frequency domain. Causality is enforced by applying a strict causal mask $\mathcal{M}(k, \omega_l)$ to the decoder transfer kernel $D_\theta\{\cdot\}$:

$$\mathbf{P}(k, \omega_l) = D_\theta\{\mathbf{U}(\leq k, \cdot)\} \odot \mathcal{M}(k, \omega_l), \tag{4}$$

where k is the short-time Fourier transform (STFT) frame index, $\mathbf{U}(\leq k, j)$ is the multi-channel structural excitation history up to frame k, and the symbol $\odot$ denotes element-wise multiplication. This ensures that the resulting time-domain mapping does not depend on future inputs and remains bounded-input bounded-output (BIBO) stable under bounded excitation. Passivity-inspired energy consistency is enforced by bounding the frequency-domain gain:

$$\sigma_{\max}(\mathbf{H}_\theta(\omega_l)) \leq \gamma(\omega_l), \tag{5}$$

where $\mathbf{H}_\theta(\omega_l)$ is the learned frequency-response matrix at ω_l, $\sigma_{\max}(\cdot)$ is the maximum singular value, and $\gamma(\omega_l)$ is the passivity-motivated spectral gain cap. In practice, the bound is implemented via phase-preserving amplitude shrinkage:

$$\mathbf{H}_\theta(\omega_l) \leftarrow \mathbf{H}_\theta(\omega_l) \cdot \min\left(1, \frac{\gamma(\omega_l)}{\sigma_{\max}(\mathbf{H}_\theta(\omega_l))}\right). \tag{6}$$

The operation rescales amplitudes while preserving phase, which is essential for coherent multi-path superposition.

3. Architectural analysis of physics-informed NN-MMFNet

3.1. Dual-stream encoder: Separation of transient and steady-state features

The network uses a time–frequency dual-stream encoder to capture transient impacts and steady resonances. As shown in Fig. 1, NN-MMFNet follows a left-to-right signal flow consisting of dual-stream encoding, causal fusion, and a physics-constrained decoder.

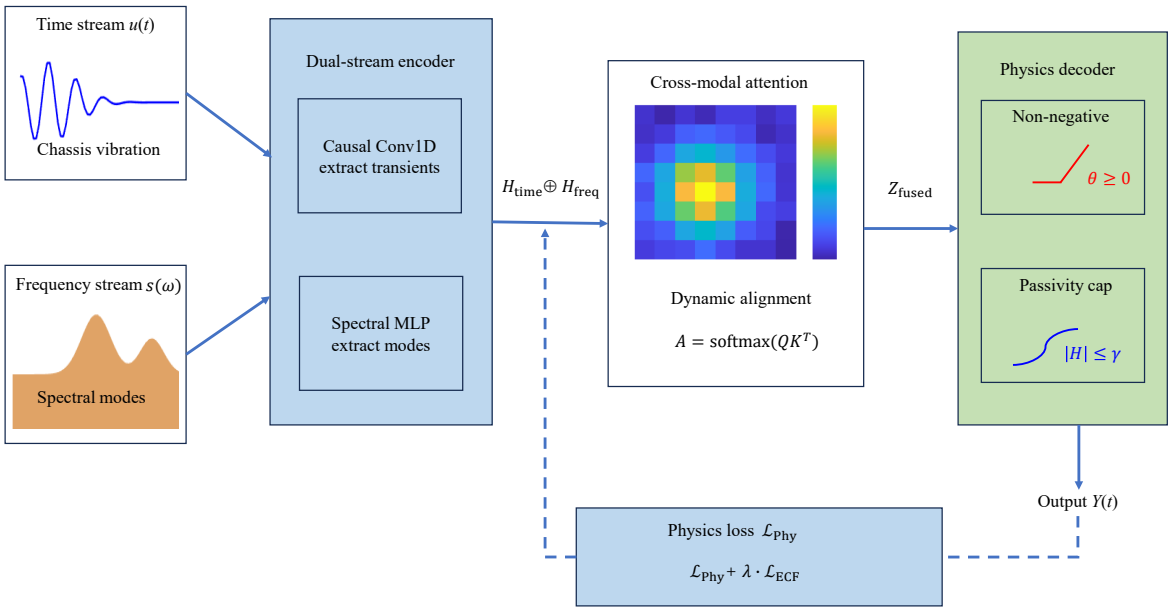

Fig. 1. Architecture of the physics-informed NN-MMFNet for BEV road noise.

The model uses a dual-stream encoder: a time stream built on 1D dilated convolutions with causal padding to capture long-range temporal dependencies without sacrificing resolution (Kiranyaz *et al.*, 2021), and a frequency stream that applies the STFT to extract 20 Hz to 300 Hz spectral features and can incorporate pre-trained noise transfer functions (NTFs) as a physics-informed prior to guide faster and more physically plausible convergence.

3.2. Cross-modal attention mechanism: Physical interpretation of energy allocation

A key element of NN-MMFNet is the cross-modal attention module. In this setting, the attention weight matrix has a direct physical interpretation: it represents the relative contribution of each excitation channel (or mode group) to the response under the current operating condition:

$$\mathbf{A}_{kl} = \text{softmax}\left(\frac{\mathbf{q}_k \cdot \mathbf{k}_l^\top}{\sqrt{d}}\right) \odot \mathbf{M}_{\text{causal}}, \tag{7}$$

where $\mathbf{q}_k$ and $\mathbf{k}_l$ are the query and key vectors, and d is the feature dimension. For interpretability and physical consistency, the attention module is constrained to be strictly causal, regularized toward a band-adjacent (near-banded) distribution to reflect continuous energy transmission, and allowed to adapt its weights under operating changes so that the effective transfer characteristics are updated online, enabling linear parameter-varying (LPV)-like behavior within a single model.

3.3. Physics-constrained decoder and non-negative gains

The decoder is designed to address the engineering need for path contribution analysis by parameterizing the transfer function in a non-negative, low-rank factorized form:

$$\mathbf{H}_\theta(\omega_l) = \underbrace{\mathbf{\Phi}_p(\omega_l) \in \mathbb{R}_{\geq 0}^{N\times R}}_{\text{mic-sideshapes}} \cdot \underbrace{\operatorname{diag}\left(\mathbf{g}(\omega_l) \in \mathbb{R}_{\geq 0}^{R}\right)}_{\frac{\text{path}}{\text{modalgains}}} \cdot \underbrace{\mathbf{\Phi}_u(\omega_l)^\top \in \mathbb{R}_{\geq 0}^{R\times M}}_{\text{excitation-sideshapes}}, \tag{8}$$

where $\mathbf{\Phi}_p(\cdot)$ and $\mathbf{\Phi}_u(\cdot)$ are microphone-side and excitation-side basis matrices, respectively, and $\mathbf{g}(\omega_l)$ collects the non-negative path gains. During training, a gradient projection algorithm $\mathit{\Pi}_{\geq 0}(\cdot)$ is used to strictly limit parameter updates within the non-negative orthant:

$$\Theta^{t+1} = \mathit{\Pi}_{\geq 0}(\Theta^t - \eta \nabla_\Theta \mathcal{L}), \tag{9}$$

where the symbol Θ denotes the constrained parameters, η is the learning rate, and $\mathcal{L}$ is the loss. This constraint enforces additive, non-negative path contributions, avoiding cancellation between positive and negative weights.

3.4. Evaluation metric system: Quantifying physical fidelity

3.4.1. Statistical accuracy metrics

Time-RMSE (RMSE_t) measures fitting precision at the waveform level [Pa]:

$$\mathrm{RMSE}_t = \sqrt{\frac{1}{T}\sum_t (p(t) - \widehat{p}(t))^2}, \tag{10}$$

where $p(t)$ and $\widehat{p}(t)$ are the measured and predicted sound pressure, and T is the number of samples.

Frequency-RMSE (RMSE_f) measures fitting precision of the spectral amplitude [Pa/Hz]. This metric is particularly important for evaluating the capture of resonance peaks.

Global RMSE is the full-band spectral error computed over the evaluation band on the same frequency grid [dB(A)]:

$$\text{Global RMSE} = \sqrt{\frac{1}{N_f}\sum_{i=1}^{N_f} (L_{\mathrm{pred}}(f_i) - L_{\mathrm{meas}}(f_i))^2}, \tag{11}$$

where $L_{\mathrm{meas}}(f_i)$ and $L_{\mathrm{pred}}(f_i)$ are the measured and predicted A-weighted sound pressure levels (SPLs) at frequency f_i, and N_f is the number of points in the 20 Hz to 300 Hz band.

3.4.2. Energy consistency factor, phase-sensitive mutual information, and causality violation ratio

Energy consistency factor (ECF) measures how well the predicted and measured spectra match in energy distribution, revealing non-physical leakage or spurious amplification:

$$\mathrm{ECF} = \frac{1}{|\Omega|}\sum_{\omega_l \in \Omega} \frac{||S_{\widehat{p}}(\omega_l)| - |S_p(\omega_l)||}{|S_p(\omega_l)| + \epsilon}, \tag{12}$$

where $S_{\widehat{p}}(\omega_l)$ and $S_p(\omega_l)$ are the predicted and measured spectra over Ω, ϵ prevents division by zero, and $\|\cdot\|$ is the Euclidean norm. Lower ECF indicates better energy consistency and reflects the effectiveness of the passivity constraint, while phase-sensitive mutual information (PSMI) assesses phase fidelity via the statistical dependence between predicted and measured phase:

$$\text{PSMI} = I_{\text{circ}}(\phi_{\widehat{p}}, \phi_p), \tag{13}$$

where $\phi_{\widehat{p}}$ and ϕ_p are the predicted and measured phase spectra, and $I_{\text{circ}}(\cdot)$ is circular mutual information. A higher PSMI indicates that the model reproduces propagation delays and phase evolution more faithfully. Causality violation ratio (CVR) measures the proportion of energy in the $t < 0$ part of the impulse response function:

$$\text{CVR} = \frac{\displaystyle\int_{-\infty}^{0} |h(t)|^2 \mathrm{d}t}{\displaystyle\int_{-\infty}^{\infty} |h(t)|^2 \,\mathrm{d}t}, \tag{14}$$

where $h(t)$ is the impulse response of the excitation-to-pressure mapping. For physical systems, CVR should be strictly zero. Any non-negligible value indicates 'pre-ringing' artifacts and causality violation.

4. Simulation data and virtual experiments

4.1. Simulation corpus and virtual testbench

A physics-based virtual testbench was built to generate paired 18-channel suspension/attachment accelerations and in-cabin pressure under controlled ISO 8608 road excitations. The model includes suspension–body–cavity coupling to reproduce structural resonances and cabin boom. Manufacturing scatter and aging were emulated by perturbing key stiffness/damping parameters within realistic bounds, forming a parametric 'virtual fleet.' The resulting frequency response function (FRF) and SPL dispersion are summarized in Fig. 2.

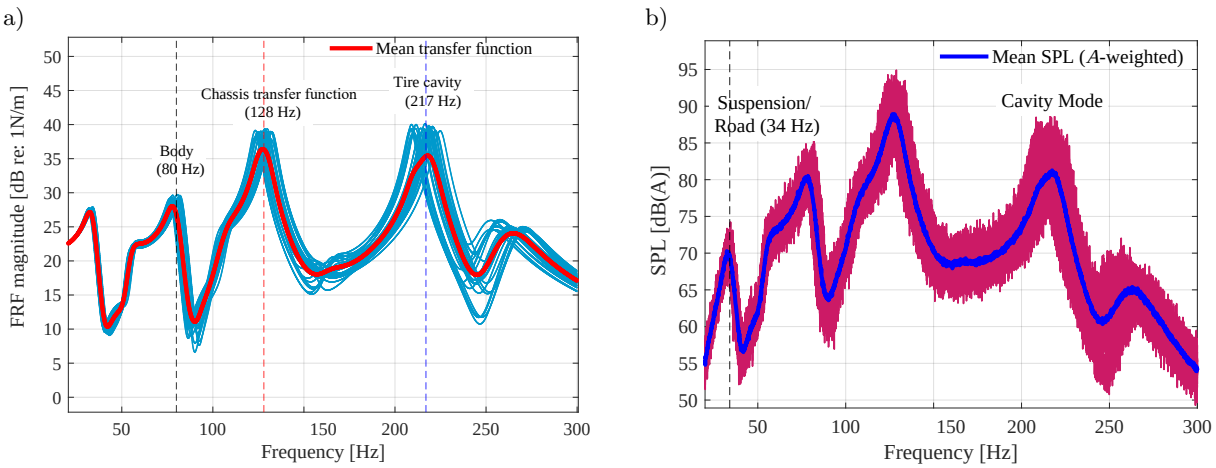

Fig. 2. Virtual fleet: parametric FRF dispersion (shaded band/area: perturbed fleet range 15 %; solid line: nominal baseline): a) parametric FRF dispersion (structure), b) in-cabin SPL dispersion (response).

Figure 3 illustrates why purely linear baselines can struggle under parameter drift and mild nonlinearities. Nonlinear coupling broadens the effective transfer behavior and can shift the modal content. By enforcing causality and passivity during learning, NN-MMFNet limits unphysical variability in the learned noise transfer function and retains the modal envelopes needed for stable boom prediction.

Evaluation protocol and baselines. Performance is reported on strictly disjoint, stratified splits (in-domain (ID), speed/road-type shifts, and a system-shift split with ±15 % plant-parameter perturbations to emulate virtual-fleet drift). Baselines are compared under matched capacity and training budgets.

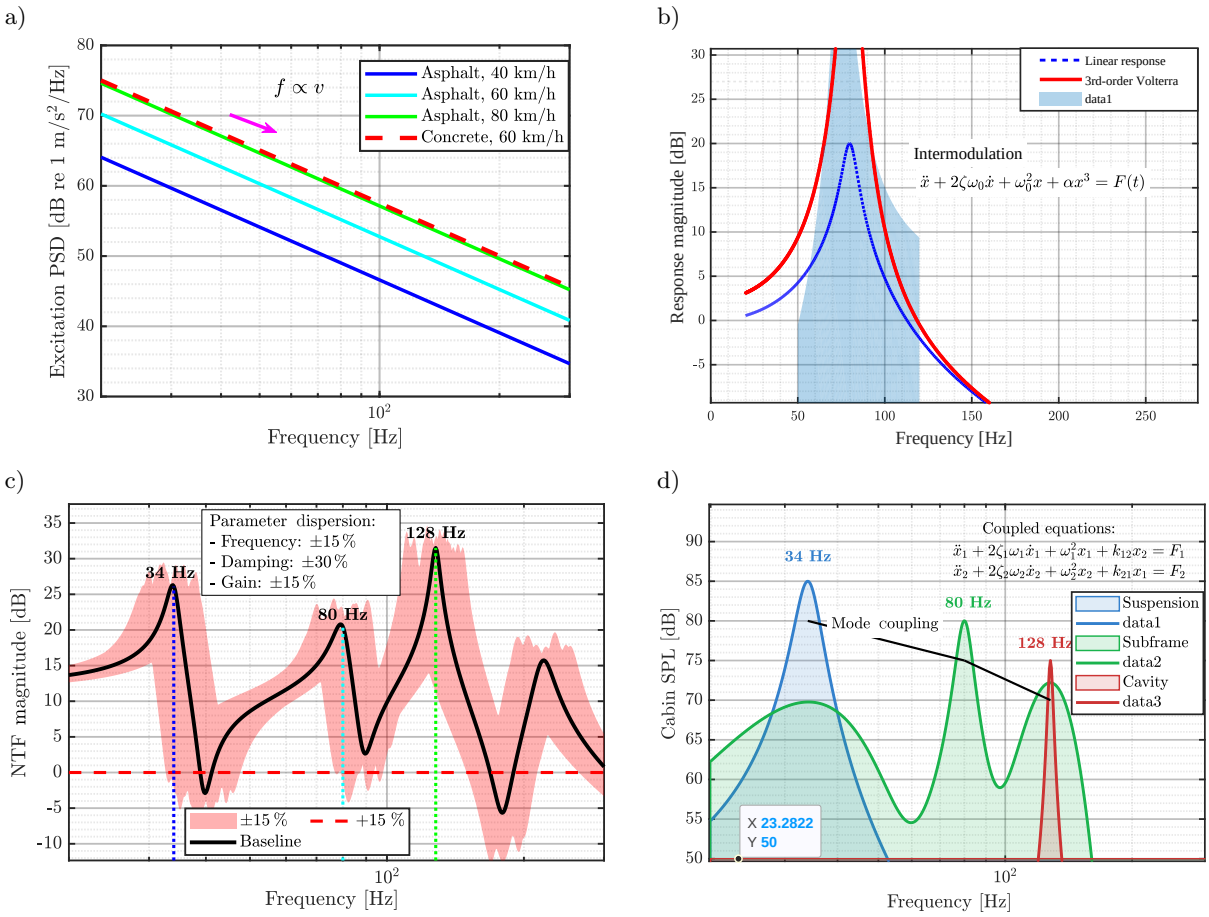

FIG. 3. Physics-constrained excitation-response analysis: a) source excitation dynamics, b) nonlinear Volterra-kernel response, c) NTF variability with physics constraints, d) key modal response envelopes.

4.2. RESULTS, ABLATIONS, AND THE SIM-TO-REAL BRIDGE

4.2.1. SIGNAL RECONSTRUCTION AND DOMAIN GENERALIZATION

NN-MMFNet provides accurate reconstruction of the coupled response in the 20 Hz to 300 Hz band. As illustrated in Fig. 4, the predicted signal follows the measured data in both time and frequency domains and resolves the main resonances at 34 Hz (suspension) and 128 Hz (cavity).

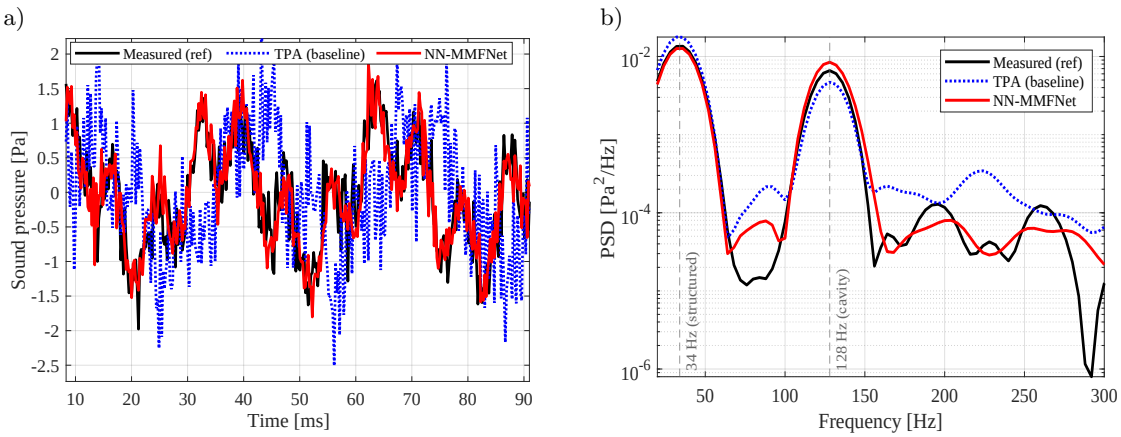

FIG. 4. Time-frequency reconstruction: a) time-domain alignment, b) spectral fidelity showing peak matching.

Unlike inverse-filtering baselines (TPA/FTM) that become ill-conditioned near resonances and amplify noise, the forward-projection decoder with a spectral cap stabilizes and effectively denoises the transfer path. The strictly causal architecture also preserves phase fidelity (phase-sensitive mutual information (PSMI) = 0.94), avoiding the non-causal phase distortions often introduced by acoustic-contribution principal component analysis (AC-PCA) or frequency-domain regularization (Cervantes-Madrid *et al.*, 2021; Shang *et al.*, 2021).

4.2.2. Statistical performance and robustness

Table 1 presents the comprehensive performance metrics. NN-MMFNet achieves the lowest RMSE and best (lowest) ECF across all splits, confirming its ability to balance numerical accuracy with physical compliance.

Table 1. Comparative performance metrics (mean and 95 % confidence intervals) across experimental splits. Lower is better ↓.

Split	Method	RMSE_t [Pa] ↓	RMSE_f [Pa/Hz] ↓	$\mathbf{\Phi}_{\mathrm{neg}}$ ↓	ECF [%] ↓
In-domain (ID)	NN-MMFNet	**29.55 [29.5, 29.6]**	**8.27**	**0.08**	**4.94**
	TPA	144.69 [144.7, 144.7]	25.78	0.26	31.15
Condition shift (CS)	NN-MMFNet	**31.47 [25.5, 37.4]**	**8.95**	**0.09**	**4.93**
	TPA	148.89 [148.7, 149.1]	27.4	0.28	29.8
System shift (SS)	NN-MMFNet	**30.85 [28.2, 33.1]**	**8.95**	**0.09**	**4.93**
	TPA	148.50 [147.1, 149.9]	27.4	0.28	29.8
Domain shift (DS)	NN-MMFNet	**22.12 [18.0, 26.9]**	**4.92**	**0.05**	**4.9**
	TPA	151.46 [150.7, 152.8]	15.7	0.16	25.72

Note: Data are derived from simulation manifests. Confidence intervals (95 % CI) are calculated via bootstrapping over segments. $\mathbf{\Phi}_{\mathrm{neq}}$ denotes the negative-contribution ratio.

The system-shift split highlights the sensitivity of conventional methods to plant drift (Fig. 5). With ±15 % parameter perturbations, TPA degrades sharply and the RMSE exceeds 148 Pa, whereas NN-MMFNet remains stable with an RMSE of 30.85 Pa, indicating an operating-adaptive mapping rather than a fixed transfer matrix. Under the unseen concrete domain, NN-MMFNet also retains low energy inconsistency with an ECF of 4.90 %, consistent with physics-informed separation of excitation $\mathbf{u}(t)$ and path dynamics $\mathbf{H}_s(\omega)$.

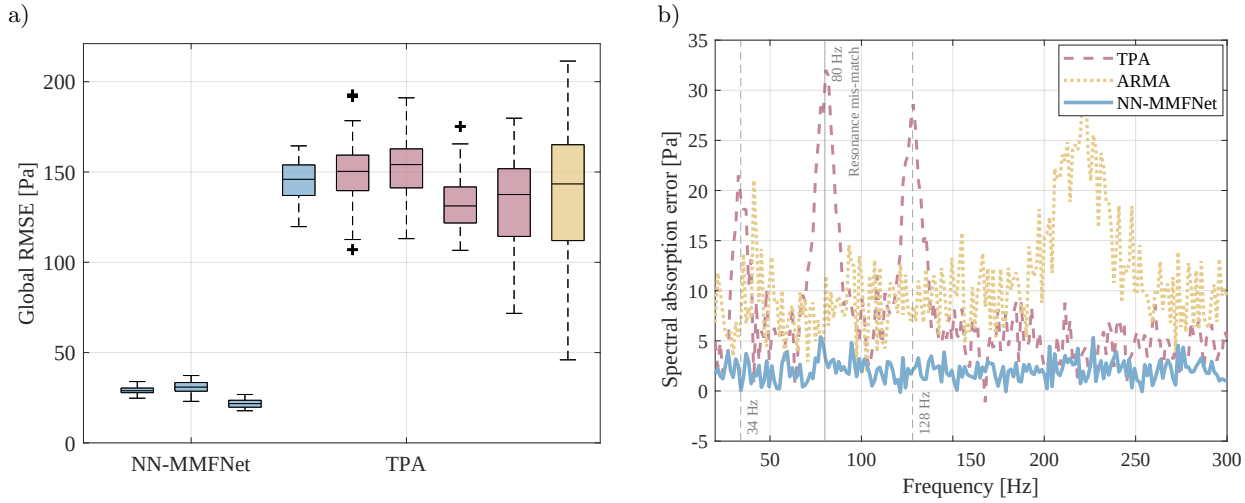

Fig. 5. Boxplots of RMSE_t distribution across in-domain, condition shift, and domain shift settings: a) statistical error distribution (global), b) frequency-resolved error profile.

4.2.3. Ablation study: Mechanism verification

Table 2 and Fig. 6 report the ablation results under the same ID split and objective. Figure 6a compares the multi-metric trade-offs, whereas Fig. 6b shows the corresponding passivity-violation rates. The full NN-MMFNet

TABLE 2. Ablation results showing the impact of removing physics constraints (ID split).

Model variant	RMSE_t [Pa] ↓	Peak-location error [Hz] ↓	PSMI ↑	ECF [%] ↓	Passivity violation [%] ↓
NN-MMFNet (full)	29.55	0	0.94	5.10	<0.01
Without spectral caps	30.12	0	0.88	**85.0**	**18.4**
Without non-negativity	28.9	0	0.42	6.12	0.02
Without normalized attention	45.3	2.5	0.81	15.12	0.05

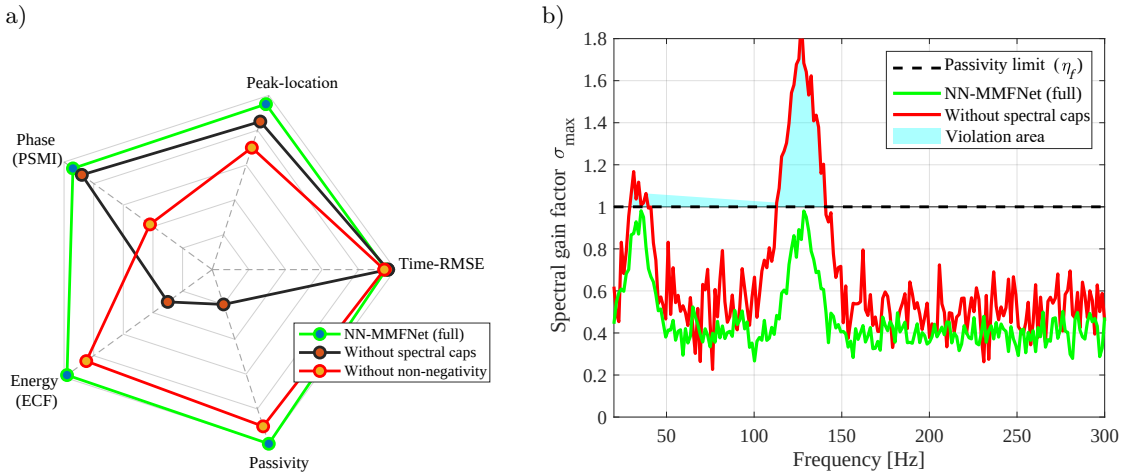

FIG. 6. Ablation study: multi-objective trade-offs and mechanism verification: a) performance trade-off (radar), b) passivity violation analysis.

achieves RMSE = 29.55 Pa, PSMI = 0.94 and ECF = 5.10 %, with passivity violations below 0.01 % and zero peak-location error. Removing the spectral cap leaves the RMSE nearly unchanged but leads to large energy inconsistency (ECF ≈ 85 %) and 18.4 % passivity violations, with non-physical amplification concentrated near 128 Hz.

Removing the non-negativity constraint slightly improves RMSE but reduces phase fidelity (PSMI = 0.42), which weakens the interpretability of the path contributions. Without normalized attention, performance degrades the most (RMSE = 45.3 Pa and peak-location error = 2.5 Hz), indicating that attention normalization is essential for robust multi-channel fusion.

5. EXPERIMENTAL VALIDATION AND ENGINEERING IMPLEMENTATION

NN-MMFNet is evaluated using a sim-to-real protocol that tests three aspects: physically constrained signal fidelity, mechanism interpretability, and robustness to parametric uncertainty.

5.1. EXPERIMENTAL SETUP AND DATA ACQUISITION PROTOCOL

Field tests were conducted on a production battery electric vehicle (SAIC-GM-Wuling F510C). Data were acquired using a 56-channel LMS SCADAS Mobile system controlled by Siemens LMS Test.Lab software. As shown in Fig. 7, the sensor layout was designed to cover the main vibro-acoustic transmission chain.

Excitation source (X): sixteen tri-axial PCB accelerometers (48 channels) were installed at key chassis hard-points. In particular, fifteen sensors monitored structure-borne inputs at suspension control arms and subframe bushings, and one sensor was mounted at the driver's seat base to capture the terminal structural transmission to the occupant.

Acoustic response (Y): four GRAS microphones were positioned at ear levels of front and rear passengers to characterize the target sound field.

Time-domain data collected at a constant speed ($v \in \{40, 60, 80\}$ km/h) were exported to MATLAB for processing and model training. To assess generalization beyond a single test vehicle, a MATLAB-based virtual fleet was generated by applying ±15 % random perturbations to the mass and stiffness matrices of a multibody

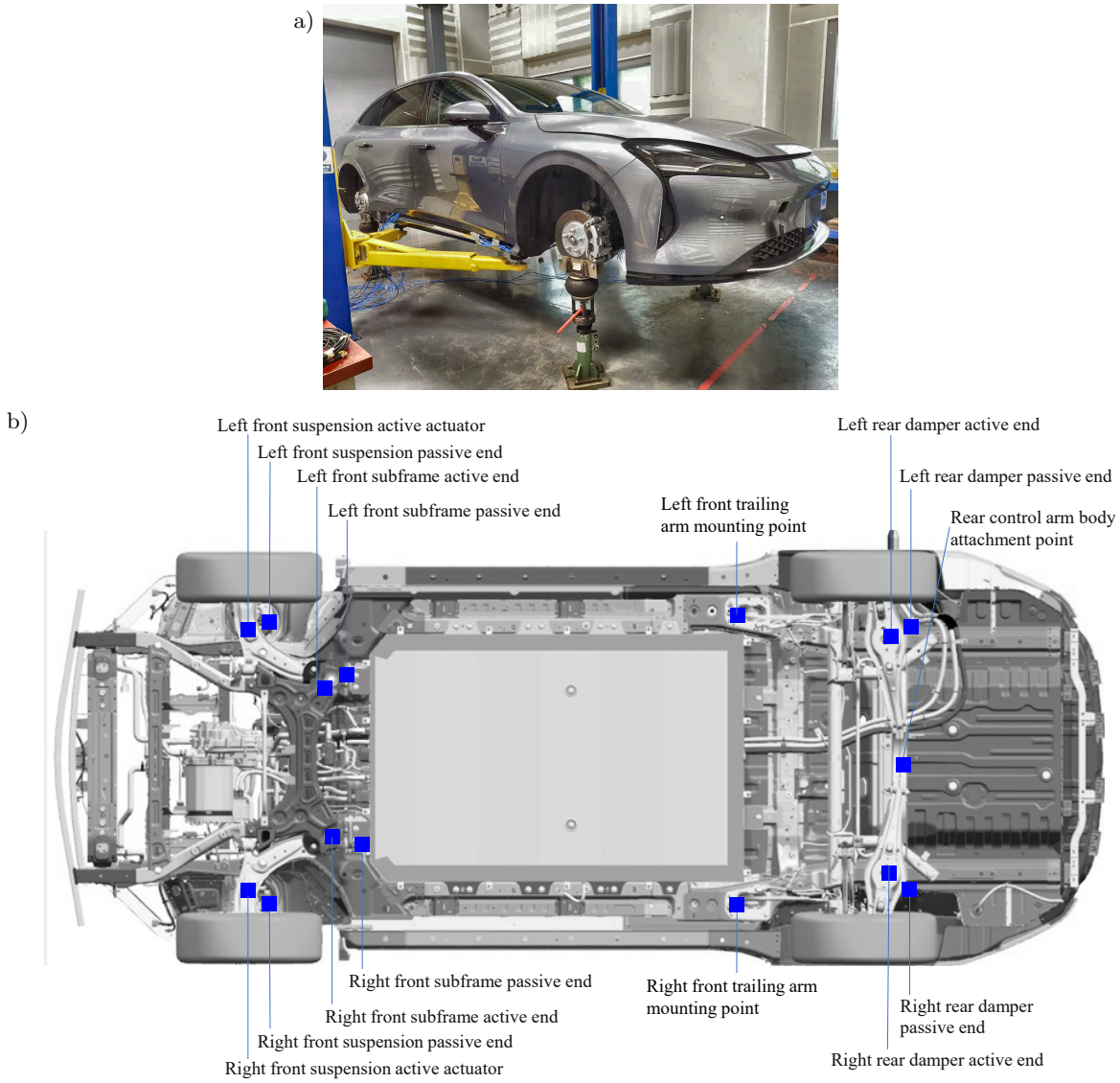

Fig. 7. Experimental test and main sensor arrangement:
a) instrumented F510C BEV platform on a four-post lift, b) topology of acceleration sensors (blue).

dynamics model, producing 100 variants. NN-MMFNet was pretrained on this virtual dataset to learn transferable coupling patterns, and then fine-tuned on the measured data.

5.2. Comparative performance analysis: Sim-to-real bridge

Pretrained on the virtual fleet, the model was fine-tuned using 15 min of real-world measurements. NN-MMFNet was benchmarked against industry-standard TPA based on matrix inversion, FTM, and ARMA baselines.

5.2.1. Frequency-domain fidelity and resonance capture

To validate the model's reconstruction capabilities under complex coupled dynamics, we analyzed the spectral fidelity and resonance-tracking accuracy within the critical 20 Hz to 300 Hz band.

Figure 8 shows full-band spectral reconstruction at 60 km/h. NN-MMFNet closely matches the measured spectrum, capturing the 128 Hz cavity boom with a 0.14 dB peak error. In contrast, TPA becomes numerically ill-conditioned above 150 Hz and exhibits ghost-energy amplification. Table 3 further confirms that NN-MMFNet reduces the global RMSE to 1.12 dB at 60 km/h, compared with 4.15 dB for TPA, demonstrating that the physics-constrained decoder suppresses phantom gain in inverse methods.

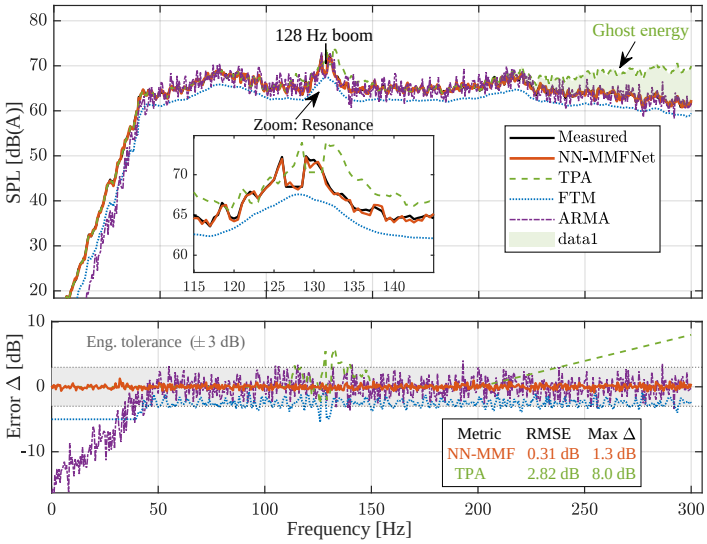

FIG. 8. Full-band spectral fidelity and error residual analysis (60 km/h).

TABLE 3. Quantitative evaluation of spectral fidelity and generalization capabilities (40 km/h to 80 km/h).

Cruising speed [km/h]	Method	Global RMSE [dB(A)] ↓	ECF [%] ↓	Suspension mode (34 Hz) [Δ dB]	Subframe mode (80 Hz) [Δ dB]	Cavity boom (128 Hz) [Δ dB]
40 (low load)	NN-MMFNet	1.10	4.65	0.08	0.12	0.15
	TPA	3.52	26.50	0.45	1.85	2.80
	FTM	3.25	24.10	1.20	2.50	1.95
	ARMA	2.80	14.50	3.50	0.90	1.20
60 (rated load)	NN-MMFNet	1.12	4.80	0.06	0.11	0.14
	TPA	4.15	31.15	0.52	2.10	3.20
	FTM	3.80	29.80	1.45	3.10	2.40
	ARMA	2.55	15.20	4.10	1.15	1.80
80 (high load)	NN-MMFNet	1.25	5.12	0.15	0.22	0.20
	TPA	6.50	42.30	0.95	3.50	**5.50**
	FTM	4.20	33.50	1.80	3.80	2.90
	ARMA	3.10	18.60	4.50	1.50	2.50

Table 3 and Fig. 9 evaluate performance from 40 km/h to 80 km/h. NN-MMFNet maintains the lowest ECF (below 5.12 %) and global RMSE (below 1.25 dB) across speeds. At 80 km/h, conventional methods degrade due to bushing nonlinearity, and TPA reaches a 5.5 dB error at 128 Hz, whereas NN-MMFNet remains stable at 0.2 dB, indicating attention-based adaptation to parametric drift and a physically consistent spectral mapping.

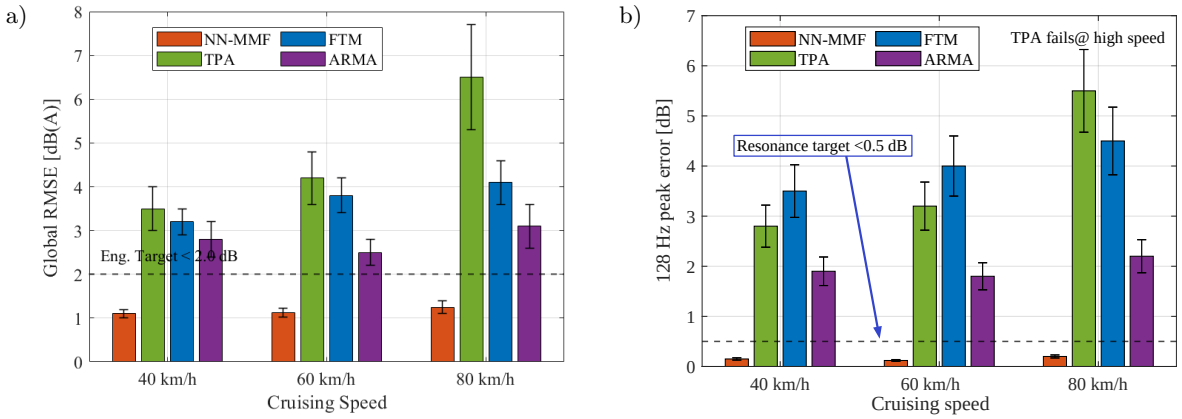

FIG. 9. Operational robustness and generalization (40 km/h to 80 km/h):
a) global accuracy across speeds, b) resonance capture fidelity (128 Hz).

5.2.2. Time-domain transient fidelity and causal verification

Physical validity under non-stationary conditions was assessed via impulse-response analysis, testing whether transient structural decay is captured while strictly preserving causality, which is often violated by block-based processing.

Figure 10 shows that TPA captures the overall decay but produces non-physical pre-ringing at negative time due to acausal processing. Table 4 quantifies this via the CVR, defined as the fraction of pre-cursor energy in the impulse response. NN-MMFNet reduces this leakage to below 0.01 %, compared with 8.40 % for TPA, confirming the effectiveness of the physics-constrained decoder.

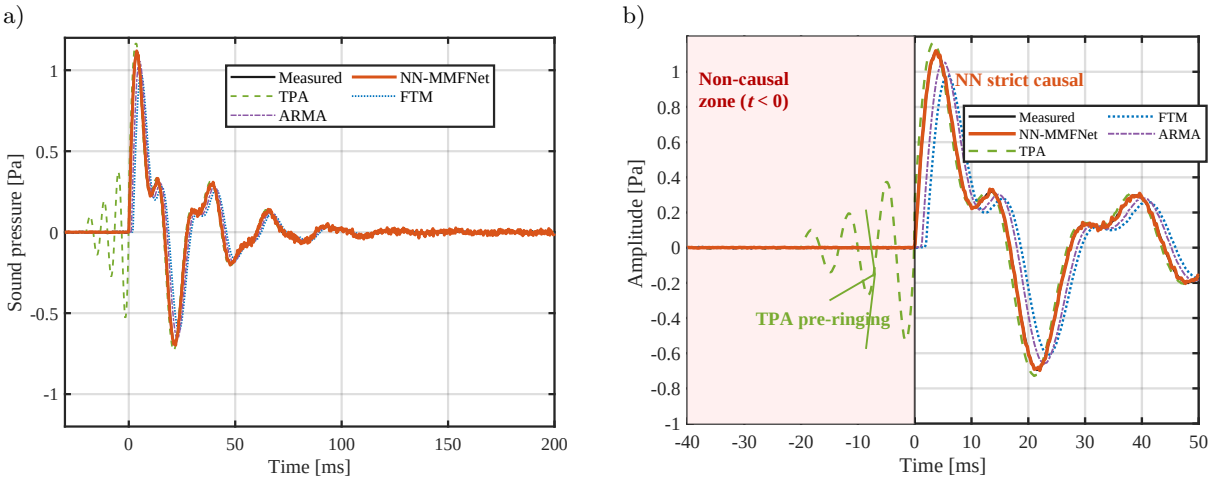

Fig. 10. Transient impulse response and causal verification:
a) global transient impulse response, b) causality check (zoom at impact).

Table 4. Quantitative evaluation of time-domain fidelity and causal consistency.

Method	RSME$_t$ [Pa] ↓	Peak phase error [°] ↓	CVR [%] ↓	Physical interpretation
NN-MMFNet	0.27	2.1	<0.01	Strictly causal
TPA (baseline)	1.23	12.5	8.40	Pre-ringing
FTM	0.61	15.0	3.20	Phase lag
ARMA	0.45	5.2	0.50	Time delay

Figure 11 shows stable error dynamics, with NN-MMFNet limiting peak phase error to 2.1° within the 5° engineering tolerance for coherent synthesis, while baselines exhibit phase drift. NN-MMFNet therefore achieves an RMSE$_t$ of 0.27 Pa.

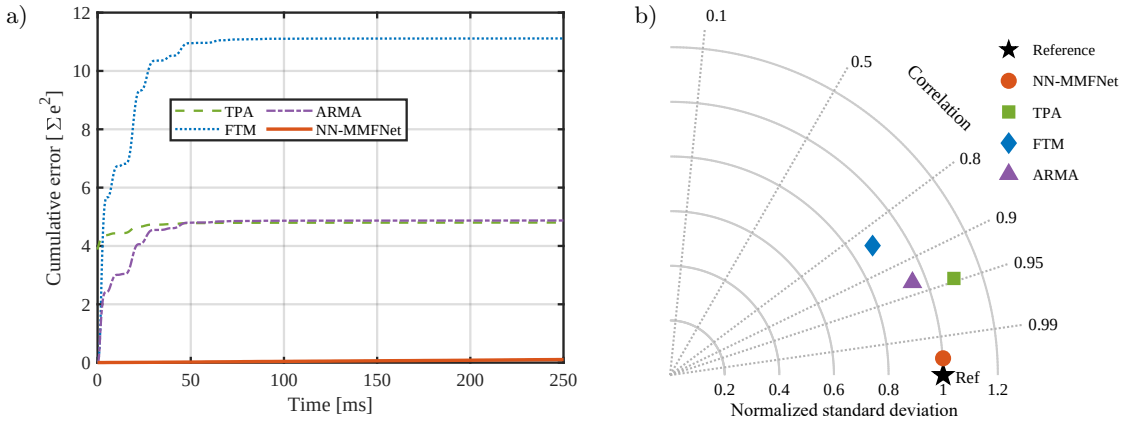

Fig. 11. Error evolution dynamics and statistical fidelity assessment:
a) error accumulation evolution, b) Taylor diagram (overall fidelity).

For the peak transient response, NN-MMFNet reduces the waveform error to 0.27 Pa, compared with 1.23 Pa for the TPA baseline. The Taylor diagram in Fig. 11b provides a consistent summary, placing NN-MMFNet closest to the reference.

5.3. Engineering application: Closed-loop diagnosis and targeted optimization

To assess practical usefulness beyond offline prediction, the non-negative attention map was used to localize the dominant structure-borne transmission path associated with the 128 Hz boom on the F510C platform. Across operating points, the attention consistently concentrated on the rear subframe-to-body mounting channel, indicating a stiff coupling in that load path.

Guided by this diagnosis, a targeted stiffness adjustment was implemented at the identified mount. Vehicle tests confirmed a 4.2 dB(A) reduction of the 128 Hz boom peak without introducing non-causal artifacts or violating passivity in the reconstructed transfer behavior. Additional multi-metric deployment analysis and secondary cases are provided in Appendix. A zero-shot cross-platform check on the F710C platform showed that the physics-constrained representation remains competitive relative to classical TPA baselines under domain shift (details in Appendix).

6. Conclusion

This paper proposed NN-MMFNet for operating-dependent structure-borne road-noise prediction in BEVs. The model integrates a dual-stream time–frequency encoder with strictly causal fusion/decoding, and passivity-enforced spectral gain control, while non-negative attention provides interpretable path contributions. With virtual-fleet pretraining followed by sim-to-real fine-tuning, the method improves robustness to speed/load variations and plant-parameter drift. Experiments on a production vehicle show a 1.12 dB(A) full-band spectral RMSE at 60 km/h, and the 128 Hz boom component is predicted within 0.14 dB(A), with a passivity-violation rate below 0.01 %. The inferred contribution map suggests the rear subframe mount as a dominant path, and a targeted stiffness update achieves a measured 4.2 dB(A) cabin-noise reduction. Future work will extend validation to broader road classes and multi-source coupling.

Appendix. Extended algorithmic benchmarking and cross-platform performance benchmarking

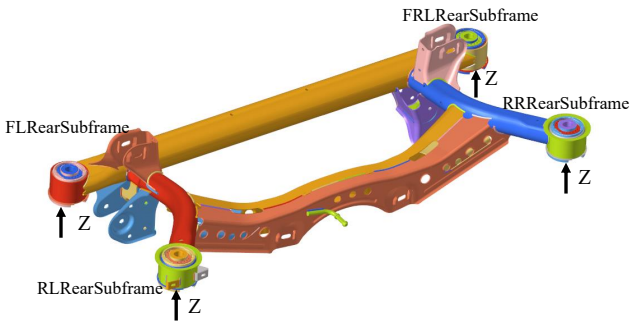

Fig. 12. Rear subframe bushing connection points.

Table 5. Quantitative performance comparison on the unseen F710C platform.

Method	Global RMSE [dB(A)] ↓	ECF [%] ↓	31 Hz error [dB] ↓	141 Hz error [dB] ↓	PSMI (phase fidelity) ↑
NN-MMFNet	0.47	6.37	0.1	0.08	0.94
TPA (matrix inv.)	3.56	107.89	5.41	6.43	0.35
FTM (regularized)	2.87	39.26	2.76	5.79	0.65
ARMA (parametric)	3.84	37.07	2.43	9.52	0.55

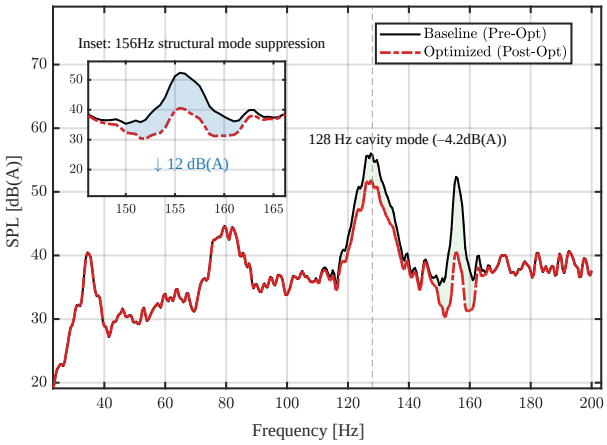

FIG. 13. F510C spectral optimization and gradient stiffness verification.

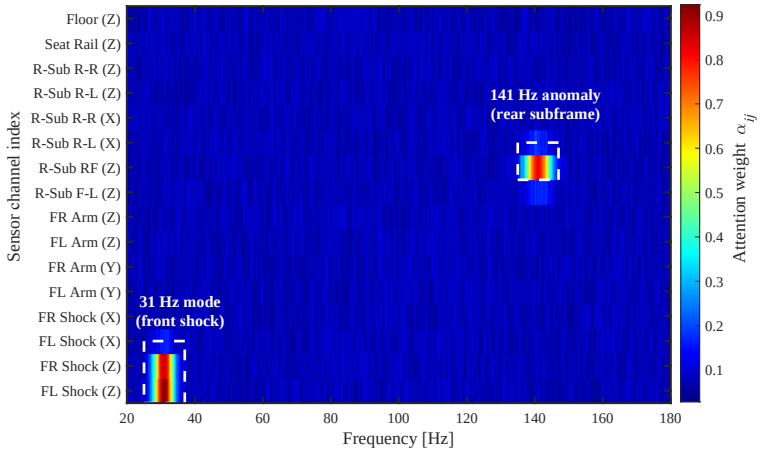

FIG. 14. Zero-shot blind diagnosis on the F710C platform: distinct pathology identification.

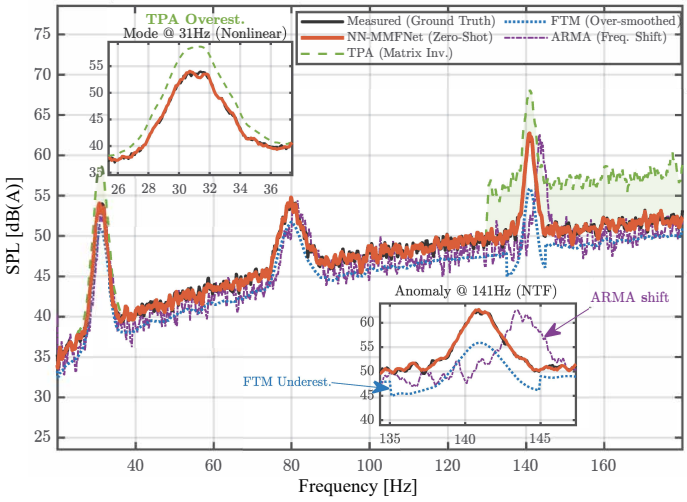

FIG. 15. Cross-platform spectral generalization performance.

FUNDINGS

This work was supported by the Guangxi Science and Technology Major Project (grant no. AA24206071) and the Guangxi Basic Ability Enhancement Program for Young and Middle-aged University Teachers (grant no. 2024KY1454).

Authors' contributions

Haijun Wang contributed to conceptualization, methodology, software development, and writing. Zengjun Lu and Xianghua He contributed to data curation, validation, review, and writing. Zhijie Huang and Tie Xu contributed to supervision and funding acquisition. All authors read and approved the final manuscript.

Conflict of interests

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Acknowledgments

We sincerely appreciate all the engineers and project participants for their invaluable contributions and support.

References

1. Cervantes-Madrid G., Peral-Orts R., Campillo-Davo N., Campello-Vincente H. (2021), Inverse transfer path analysis, a different approach to shorten time in NVH assessments, *Applied Acoustics*, **181**: 108178, https://doi.org/10.1016/j.apacoust.2021.108178.
2. Cheng W. *et al.* (2022), AR model-based crosstalk cancellation method for operational transfer path analysis, *Journal of Mechanical Science and Technology*, **36**: 1131–1144, https://doi.org/10.1007/s12206-022-0206-7.
3. Cheng W., Chu Y., Chen X., Zhou G., Blamaud D., Lu J. (2020), Operational transfer path analysis with crosstalk cancellation using independent component analysis, *Journal of Sound and Vibration*, **473**: 115224, https://doi.org/10.1016/j.jsv.2020.115224.
4. Cheng W., Lu Y., Zhang Z. (2016), Tikhonov regularization-based operational transfer path analysis, *Mechanical Systems and Signal Processing*, **75**: 494–514, https://doi.org/10.1016/j.ymssp.2015.12.025.
5. de Klerk D., Ossipov A. (2010), Operational transfer path analysis: Theory, guidelines and tire road noise application, *Mechanical Systems and Signal Processing*, **24**(7): 1950–1962, https://doi.org/10.1016/j.ymssp.2010.05.009.
6. de Klerk D., Rixen D.J. (2010), Component transfer path analysis method with compensation for test bench dynamics, *Mechanical Systems and Signal Processing*, **24**(6): 1693–1710, https://doi.org/10.1016/j.ymssp.2010.01.006.
7. Gao L. *et al.* (2024), Operational transfer path analysis with crosstalk cancellation based on least variance spectrum estimation, *Journal of Mechanical Science and Technology*, **38**: 5311–5322, https://doi.org/10.1007/s12206-024-0907-1.
8. Gustavsen B., Semlyen A. (1999), Rational approximation of frequency domain responses by vector fitting, *IEEE Transactions on Power Delivery*, **14**(3): 1052–1061, https://doi.org/10.1109/61.772353.
9. Gustavsen B., Semlyen A. (2001), Enforcing passivity for admittance matrices approximated by rational functions, *IEEE Transactions on Power Systems*, **16**(1): 97–104, https://doi.org/10.1109/59.910786.
10. Huang H., Lim T.C., Wu J., Ding W., Pang J. (2023), Multitarget prediction and optimization of pure electric vehicle tire/road airborne noise sound quality based on a knowledge- and data-driven method, *Mechanical Systems and Signal Processing*, **197**: 110361, https://doi.org/10.1016/j.ymssp.2023.110361.
11. Jia X., Zhou L., Huang H., Pang J., Yang L. (2024), Improving electric vehicle structural-borne noise based on convolutional neural network-support vector regression, *Electronics*, **13**(1): 113, https://doi.org/10.3390/electronics13010113.
12. Khan D., Burdzik R. (2023), Measurement and analysis of transport noise and vibration: A review of techniques, case studies, and future directions, *Measurement*, **220**: 113354, https://doi.org/10.1016/j.measurement.2023.113354.
13. Kiranyaz S., Avci O., Abdeljaber O., Ince T., Gabbouj M., Inman D.J. (2021), 1D convolutional neural networks and applications: A survey, *Mechanical Systems and Signal Processing*, **151**: 107398, https://doi.org/10.1016/j.ymssp.2020.107398.

14. Kong L. *et al.* (2025), A novel operational transfer path analysis based on the complex-valued crosstalk elimination method, *Measurement*, **253**(Part D): 117813, https://doi.org/10.1016/j.measurement.2025.117813.

15. Ma Y., Dai R., Liu T., Liu J., Yang S., Wang J. (2025), Research on vehicle road noise prediction based on AFW-LSTM, *Machines*, **13**(5): 425, https://doi.org/10.3390/machines13050425.

16. Masri J., Amer M., Salman S., Ismail M., Elsisi M. (2024), A survey of modern vehicle noise, vibration, and harshness: A state-of-the-art, *Ain Shams Engineering Journal*, **15**: 102957, https://doi.org/10.1016/j.asej.2024.102957.

17. Mohammadi N. (2023), Airborne and structure-borne noise control in the MB truck cabin interior by the noise reduction in the transmission path, *Archives of Acoustics*, **48**(1): 93–101, https://doi.org/10.24425/aoa.2022.142908.

18. Moorhouse A.T., Elliott A.S., Evans T.A. (2009), In situ measurement of the blocked force of structure-borne sound sources, *Journal of Sound and Vibration*, **325**(4–5): 679–685, https://doi.org/10.1016/j.jsv.2009.04.035.

19. Münder M., Carbon C.C. (2022), A literature review [2000–2022] on vehicle acoustics: Investigations on perceptual parameters of interior soundscapes in electrified vehicles, *Frontiers in Mechanical Engineering*, **8**: 974464, https://doi.org/10.3389/fmech.2022.974464.

20. Ortega Almirón J., Bianciardi F., Corbeels P., Pieroni N., Kindt P., Desmet W. (2022), Vehicle road noise prediction using component-based transfer path analysis from tire test-rig measurements on a rolling tire, *Journal of Sound and Vibration*, **523**: 116694, https://doi.org/10.1016/j.jsv.2021.116694.

21. Park U., Kang Y.J. (2024), Operational transfer path analysis based on neural network, *Journal of Sound and Vibration*, **579**: 118364, https://doi.org/10.1016/j.jsv.2024.118364.

22. Raissi M., Perdikaris P., Karniadakis G.E. (2019), Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, *Journal of Computational Physics*, **378**: 686–707, https://doi.org/10.1016/j.jcp.2018.10.045.

23. Shang Z., Hu F., Zeng F., Wei L., Xu Q., Wang J. (2021), Research of transfer path analysis based on contribution factor of sound quality, *Applied Acoustics*, **173**: 107693, https://doi.org/10.1016/j.apacoust.2020.107693.

24. van der Seijs M.V., de Klerk D., Rixen D.J. (2016), General framework for transfer path analysis: History, theory and classification of techniques, *Mechanical Systems and Signal Processing*, **68–69**: 217–244, https://doi.org/10.1016/j.ymssp.2015.08.004.

25. Yang M., Dai P., Yin Y., Wang D., Wang Y., Huang H. (2025), Predicting and optimizing pure electric vehicle road noise via a locality-sensitive hashing transformer and interval analysis, *ISA Transactions*, **157**: 556–572, https://doi.org/10.1016/j.isatra.2024.11.059.

26. Zhang E., Chen Y., Su L., Zhonglian R., Chen X., Jiang S. (2024), Evaluation modeling of electric bus interior sound quality based on two improved XGBoost algorithms using GS and PSO, *Archives of Acoustics*, **49**(3): 307–317, https://doi.org/10.24425/aoa.2024.148794.

27. Zhu H., Zhao J., Wang Y., Ding W., Pang J., Huang H. (2024), Improving of pure electric vehicle sound and vibration comfort using a multi-task learning with task-dependent weighting method, *Measurement*, **233**: 114752, https://doi.org/10.1016/j.measurement.2024.114752.

Archives of Acoustics Vol. 51, No. 2, pp. 283–292 (2026), https://doi.org/10.24423/archacoust.2026.4407

Research Paper

A Study of Damage Mode Recognition of Polypropylene Fiber-Reinforced Recycled Aggregate Concrete Based on Principal Components of Acoustic Emission Signals

Qianxu Chen[(1)], Xin Yang[(2)*]

[(1)] *James Watt School of Engineering, University of Glasgow*
Glasgow, United Kingdom

[(2)] *Institute of Marine and Environmental Geotechnical Engineering, Fujian University of Technology*
Fuzhou, China

*Corresponding Author: yangxin546@163.com

Received January 4, 2026; revised March 19, 2026; accepted March 20, 2026; available online March 25, 2026; version of record May 27, 2026; published issue June 24, 2026.

To investigate the principal components of acoustic emission (AE) signals and the damage modes of polypropylene fiber (PPF)-reinforced recycled concrete, ten groups of specimens with coarse aggregate (CA) replacement rates of 0 % and 25 % and with different particle sizes, are designed and fabricated. Uniaxial compression AE tests are conducted to obtain AE parameters during the fracture process of PPF-reinforced recycled concrete. In this study, the Pearson correlation coefficient is employed to investigate the correlations among AE parameters. Then, principal component analysis (PCA) is performed on the AE signals to conduct dimensionality reduction of the multi-dimensional data. On this basis, the optimal number of clusters for the principal components of AE signals is determined based on the silhouette coefficient. Finally, the k-means clustering algorithm is introduced to perform cluster analysis on the principal components of AE signals of PPF-reinforced recycled concrete. The clustering results are compared with each other to explore the characteristics of each cluster and to identify the corresponding damage mode for each cluster. The discriminability of AE parameters with respect to damage modes is also investigated. The research findings can provide a reference for predicting the fracture mechanism of PPF-reinforced recycled concrete.

Keywords: polypropylene fiber (PPF)-reinforced recycled concrete, acoustic emission, Pearson correlation coefficient, principal component analysis, k-means clustering.

1. Introduction

At present, the factors restricting the widespread application of recycled concrete lie in the fact that its mechanical properties, such as compressive strength and tensile strength, are inferior to those of natural concrete. To address this problem, methods such as the addition of external fibers can be adopted. Adding polypropylene fiber (PPF) to recycled concrete is a relatively common method (Wang *et al.*, 2022).

In recent years, numerous scholars have investigated the characteristics of acoustic emission (AE) signals in fiber-reinforced concrete, including principal component analysis (PCA) (Tayfur *et al.*, 2018), digital image technology (Ashraf, Rucka, 2024; Sagar *et al.*, 2025), the b-value method (Ashraf, Rucka, 2023), 3D printing technology (Ingle, Prem, 2025), shear models (Kantekin, Bakir, 2025), fatigue behavior (Džolan *et al.*, 2024), three-point bending test (Umar *et al.*, 2023), and durability (Chkhachirou *et al.*, 2025). Zheng *et al.* (2022) distinguished two types of acoustic events, matrix cracking and steel fiber vibration, according to AE

technique. Zaki *et al.* (2023) utilized rise time/amplitude analysis to classify the damage modes. Kouta *et al.* (2021) found that both AE activity and fracture energy increase with the rise in fiber content and fiber length. Saha and Sagar (2021) classified the AE signals generated by fiber-reinforced concrete into two categories via machine learning methods: cement matrix cracking and fiber pull-out, and pointed out that the classification of AE waveforms might facilitate the understanding of damage evolution during the fracture process.

In the current paper, AE parameters are obtained by conducting AE monitoring tests on PPF-reinforced recycled concrete. Based on methods including the Pearson correlation coefficient, PCA, and k-means clustering, an analysis is conducted on the principal components of the AE parameters and the corresponding damage modes of PPF-reinforced recycled concrete. The research findings can provide a reference for predicting the fracture mechanism of PPF-reinforced recycled concrete.

2. Experimental details

2.1. Sample preparation

The PPFs employed in this experiment were manufactured by Shandong Runlin Wood Industry Co., Ltd., and their main properties are presented in Table 1.

Table 1. Physical and mechanical parameters of PPFs.

Fiber type	Diameter [mm]	Length [mm]	Tensile strength [MPa]	Fracture strength [MPa]	Elongation at break [%]	Initial modulus [GPa]	Density [g/cm^3]	Recommended dosage [kg/m^3]
Micro-PPF	0.036	19	≥450	450	17	4.8	0.91	0.9
Macro-PPF1	0.9	30	≥550	500	24	6.6	0.91	6.0
Macro-PPF2	0.9	50	≥550	500	24	6.6	0.91	6.0

In this experiment, ordinary Portland cement is used as the cementitious material, while medium sand from Zone II is adopted as the fine aggregate, with a particle size range of 0.15 mm to 4.75 mm. The particle size range of the coarse aggregate (CA) is 5 mm to 20 mm. The recycled CA is derived from waste C30 concrete, which is subjected to impurity removal, crushing, and subsequent sieving to obtain the recycled CA meeting the experimental requirements, as illustrated in Fig. 1.

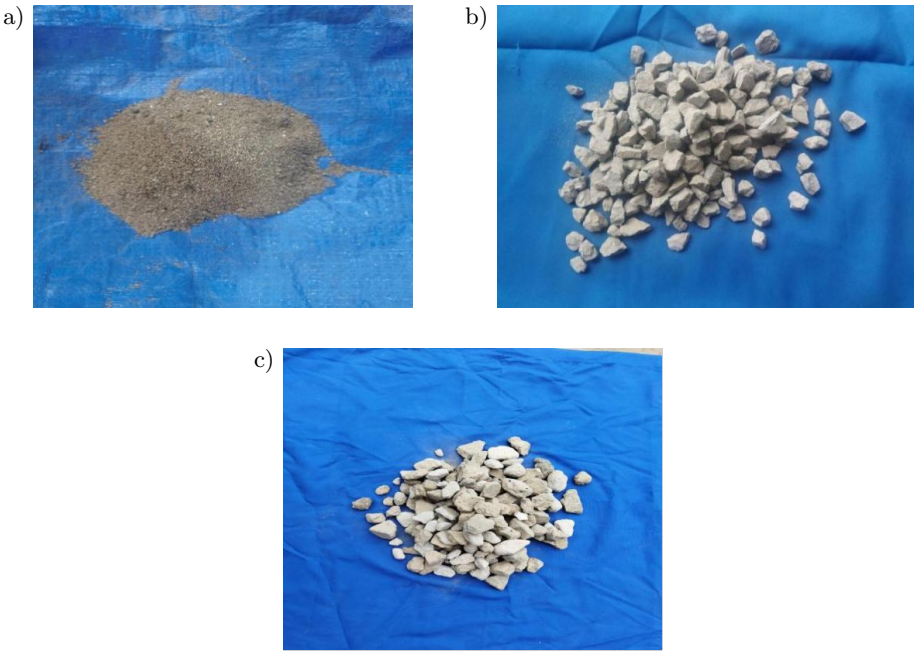

Fig. 1. Fine aggregate (a), natural coarse aggregate (b), and recycled coarse aggregate (c).

The concrete is prepared with a strength grade of C30, with a total of ten groups of specimens fabricated. The mix proportions are as follows: 358 kg/m^3 of cement, 706.15 kg/m^3 of medium sand, 1120.85 kg/m^3 of CA, and 215 kg/m^3 of water. The water-cement ratio is kept constant at 0.6 for all groups of specimens, while the variables are the CA replacement rate and fiber content. Specifically, the replacement rates of recycled CA are set at 0 % and 25 %, labeled as R-0 and R-25, respectively. In this experiment, both natural CA and recycled CA have a particle size range of 5 mm to 20 mm, with the proportion of the 5 mm to 10 mm fraction to the 10 mm to 20 mm fraction being 1:1. Specimen R-X-1 is a plain concrete specimen with no PPF added. Specimen R-X-2 incorporates microfibers at a dosage of 0.9 kg/m^3, while specimen R-X-3 incorporates macrofibers at a dosage of 6 kg/m^3. In addition, specimens R-X-4 and R-X-5 incorporate a hybrid blend of macrofibers and microfibers, with the total fiber dosage maintained at 6 kg/m^3. The specific mix proportions are provided in Table 2. Each group tested once for PCA.

Table 2. Mix proportions and properties of their test specimens.

Specimen no.	CA [kg/m^3]		Fiber length and fiber diameter [mm]	Fiber dosage [kg/m^3]	Compressive strength [MPa]
	Natural 5–10 / 10–20 [mm]	Recycled 5–10 / 10–20 [mm]			
Zero CA substitution					
R-0-1	560/560	0	None	0	37.68
R-0-2	560/560	0	19/0.036	0.9	38.99
R-0-3	560/560	0	50/0.9	6	45.94
R-0-4	560/560	0	19/0.036+30/0.9	0.9+5.1	42.92
R-0-5	560/560	0	19/0.036+50/0.9	0.9+5.1	41.13
25 % CA substitution					
R-25-1	420/420	140/140	None	0	33.23
R-25-2	420/420	140/140	19/0.036	0.9	35.56
R-25-3	420/420	140/140	50/0.9	6	44.09
R-25-4	420/420	140/140	19/0.036+30/0.9	0.9+5.1	43.46
R-25-5	420/420	140/140	19/0.036+50/0.9	0.9+5.1	51.75

2.2. Tests

In accordance with the requirements of Ministry of Housing and Urban-Rural Development of the PRC (2019), cube specimens with a side length of 150 mm are prepared. After the concrete mixture is poured into the molds, the specimens are cured at room temperature for one day. Subsequently, the specimens are demolded, labeled, and then subjected to natural curing in a curing room for 28 days. An HCT306B compression testing machine (Fig. 2) is employed for the uniaxial compression test, with a loading rate of 0.5 MPa/s. An AMSY-6 AE testing

Fig. 2. HCT306B compression testing machine.

system is adopted for AE monitoring, and the threshold of the AE instrument is set at 40 dB, with a sampling frequency of 5 MHz, to minimize the impact of ambient noise during the test.

3. Damage mode identification based on AE parameters

3.1. Preliminary screening of AE parameters based on the Pearson correlation coefficient method

The Pearson correlation coefficient can reflect the degree of linear correlation between two variables; its value ranges from −1 to 1, with a larger absolute value indicating a stronger correlation. Therefore, the Pearson correlation coefficient method can be adopted for the preliminary screening of AE characteristic parameters, with the aim of selecting parameters with low correlation coefficients and high mutual independence as far as possible, thereby reducing the size of the characteristic parameter set. Equation (1) presents the calculation formula for the Pearson correlation coefficient r. When $r = 1$, it indicates a perfect positive linear correlation between the two variables, when $r = -1$, it indicates a perfect negative linear correlation, and when $r = 0$, it indicates no linear correlation between them:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{1}$$

where x_i and y_i represent the i-th observed values of the two variables x and y, respectively, $\overline{x}$ and $\overline{y}$ denote the sample means of variables x and y, and n is the sample size.

In this paper, six AE characteristic parameters, namely amplitude, rise time, duration, count, energy, and dominant frequency, were selected for the Pearson correlation coefficient analysis. These parameters characterize the AE signals in multiple aspects in both the time and frequency domains, and are thus capable of comprehensively reflecting the AE behavior during the damage process.

Table 3 presents the Pearson correlation coefficients among the six AE characteristic parameters, namely amplitude, rise time, duration, count, energy, and dominant frequency. Based on the values of the correlation coefficients, parameters with a coefficient greater than 0.6 were regarded as having a strong correlation, in which case one of the characteristic parameters could be discarded. As can be seen from Table 3, the Pearson correlation coefficients between the three parameters (rise time, energy, and dominant frequency) and the other parameters are mostly less than 0.6, indicating low correlation and high independence of these three parameters. The Pearson correlation coefficients among the three parameter pairs – amplitude and count, amplitude and duration, and count and duration – are mostly greater than 0.6, indicating a strong correlation between each pair. Among these parameters, amplitude and count have well-defined physical meanings: the higher their values, the stronger the acoustic emission activity. In contrast, duration generally needs to be interpreted comprehensively in combination with other parameters. Therefore, five AE characteristic parameters, namely amplitude, rise time, count, energy and dominant frequency, were selected for PCA.

3.2. Selection of characteristic parameters based on PCA

PCA is a data dimensionality reduction method that transforms standardized parameter data into several uncorrelated principal component variables via dimensionality reduction, thus capturing most of the information contained in the original dataset. To eliminate the influence of dimensional differences among different characteristic parameters, the data need to be standardized prior to performing PCA.

Assuming that the original dataset is denoted as X, it is then subjected to standardization to yield the sample matrix X_{n*p}, where n is the number of samples and p is the number of features. The covariance matrix R_{p*p} is

Table 3. Pearson correlation coefficients between different AE parameters.

Specimen no.	AE parameter	Amplitude	Rise time	Duration	Count	Energy
R-0-1	Rise time	0.43	–	–	–	–
	Duration	0.75	0.63	–	–	–
	Count	0.87	0.5	0.84	–	–
	Energy	0.53	0.13	0.26	0.51	–
	Dominant frequency	0.17	0.1	0.16	0.22	0.1
R-0-2	Rise time	0.63	–	–	–	–
	Duration	0.84	0.75	–	–	–
	Count	0.87	0.7	0.94	–	–
	Energy	0.55	0.31	0.39	0.51	–
	Dominant frequency	−0.05	−0.1	−0.11	−0.02	0.03
R-0-3	Rise time	0.19	–	–	–	–
	Duration	0.57	0.36	–	–	–
	Count	0.71	0.27	0.75	–	–
	Energy	0.5	−0.01	0.13	0.37	–
	Dominant frequency	0.21	0.03	0.14	0.29	0.1
R-0-4	Rise time	0.06	–	–	–	–
	Duration	0.4	0.21	–	–	–
	Count	0.49	0.08	0.58	–	–
	Energy	0.38	−0.03	0.05	0.23	–
	Dominant frequency	−0.01	−0.01	0.1	0.27	−0.03
R-0-5	Rise time	0.08	–	–	–	–
	Duration	0.34	0.17	–	–	–
	Count	0.77	0.09	0.38	–	–
	Energy	0.32	0.02	0.05	0.15	–
	Dominant frequency	0.31	0.01	0.06	0.51	0.02
R-25-1	Rise time	0.03	–	–	–	–
	Duration	0.17	0.09	–	–	–
	Count	0.58	0.02	0.3	–	–
	Energy	0.66	0.02	0.04	0.51	–
	Dominant frequency	0.08	−0.05	0.06	0.27	0.07
R-25-2	Rise time	0.04	–	–	–	–
	Duration	0.42	0.18	–	–	–
	Count	0.49	0.07	0.6	–	–
	Energy	0.48	−0.02	0.07	0.17	–
	Dominant frequency	0.01	0.002	0.07	0.22	−0.06
R-25-3	Rise time	0.35				
	Duration	0.67	0.49	–	–	–
	Count	0.75	0.41	0.83	–	–
	Energy	0.47	0.08	0.19	0.42	–
	Dominant frequency	0.005	0.01	−0.02	−0.01	0.07
R-25-4	Rise time	0.11	–	–	–	–
	Duration	0.47	0.21	–	–	–
	Count	0.48	0.11	0.64	–	–
	Energy	0.49	0.001	0.09	0.11	–
	Dominant frequency	−0.09	−0.1	0.02	0.1	−0.17
R-25-5	Rise time	0.52	–	–	–	–
	Duration	0.78	0.69	–	–	–
	Count	0.8	0.65	0.93	–	–
	Energy	0.47	0.21	0.29	0.45	–
	Dominant frequency	−0.13	−0.12	−0.18	−0.13	−0.03

calculated according to Eq. (2). The covariance matrix is then subjected to eigenvalue decomposition to derive p eigenvalues, which are then sorted in descending order to obtain $\lambda_1, \lambda_2, ..., \lambda_p$ and their corresponding

$T_1, T_2, ..., T_p$. At this point, the i-th calculated principal component can be expressed by Eq. (3), while the contribution rate of the principal component ϕ_k can be expressed by Eq. (4):

$$R_{p*p} = \frac{1}{n-1} X^T_{n*p} X_{n*p}, \tag{2}$$

$$\mathrm{PCA}(i) = T_{1i}X_1 + T_{2i}X_2 + \cdots + T_{pi}X_p, \tag{3}$$

$$\phi_k = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}, \tag{4}$$

where ϕ_k denotes the contribution rate of the k-th principal component.

Based on the results of the Pearson correlation coefficient analysis, five characteristic parameters, namely amplitude, rise time, count, energy, and dominant frequency, were selected for PCA. Table 4 presents the contribution rates of each principal component for the specimens. As can be seen from Table 4, the cumulative contribution rate of the first three principal components exceeds 75 %, indicating that the selection of these three principal components can well retain most of the information from the original dataset.

Table 4. Contribution rates of each principal component and the cumulative contribution rate of the first three principal components (PCs) [%].

Specimen no.	1	2	3	4	5	Cumulative contribution of PCs 1:3
R-0-2	56.4	20.4	13.7	6.9	2.6	90.5
R-0-3	44.7	20.5	18.5	10.9	5.4	83.7
R-0-4	35.4	22.2	20.2	13.6	8.6	77.8
R-0-5	48.8	20.5	16.6	11.1	3	85.9
R-25-2	35.9	22.7	20	13.8	7.6	78.6
R-25-3	46.2	20.4	18.2	10.4	4.8	84.8
R-25-4	35.7	23.1	19.5	14.5	7.2	78.3
R-25-5	52.6	19.8	15.8	8.3	3.5	88.2

3.3. Determination of the optimal number of clusters

In the process of conducting cluster analysis, determining a reasonable number of clusters is a key step in ensuring the reliability of the results. An excessively small number of clusters may result in different damage modes being incorrectly classified into the same category. Conversely, an excessively large number of clusters may lead to the over-segmentation of the same damage mode. In this paper, the silhouette coefficient is adopted to evaluate the number of clusters. The range of the number of clusters was initially set to 2–6, and the silhouette coefficient was used to evaluate the performance of each cluster number, thereby determining the optimal number of clusters ultimately.

The silhouette coefficient (SI) is an index that calculates two metrics: the average distance between a sample and other samples within its own cluster, and the average distance between the sample and the samples within the nearest neighboring cluster. It evaluates the compactness and separation of clustering by measuring the difference between these two distances. A larger SI indicates more reasonable sample clustering and higher clustering quality. The calculation equation for the SI is given by

$$\mathrm{SI} = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{5}$$

where $a(i)$ and $b(i)$ denote the average distance from sample i to other samples within the same cluster and to all samples in the nearest neighboring cluster, respectively.

Figure 3 presents the number of clusters calculated based on the SI index. It can be seen that the SI index reaches its maximum value when the number of clusters is set to 4; therefore, the optimal number of clusters is determined to be 4.

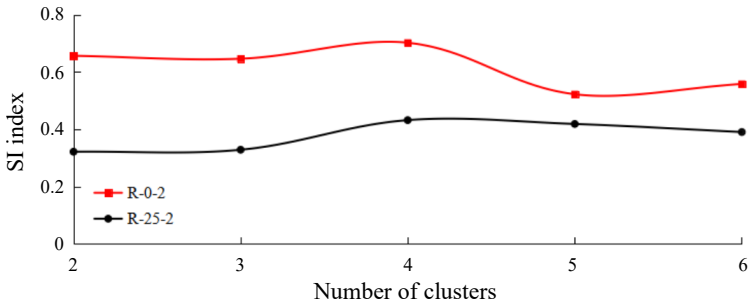

FIG. 3. Number of clusters assessed by the SI index.

4. Analysis of clustering results

In this paper, the k-means clustering algorithm was adopted to perform cluster analysis on the results of principal component analysis of AE parameters of PPF-reinforced recycled concrete, and the number of clusters was set to 4, as determined in Subsec. 3.3.

4.1. k-means clustering

k-means is an iterative clustering algorithm whose operating principle can be outlined as follows: given a sample dataset composed of n samples $X_1, X_2, ..., X_n$, the algorithm aims to partition these samples into k-distinct clusters. Here, m_i the centroid (mean vector) of the samples assigned to the i-th cluster. The algorithm employs the distanceas the distance metric. The detailed steps are as follows (Wen, 2025):

1. Initialization: randomly and uniformly select k observation samples as the initial cluster centers m_1–m_k.
2. Assignment: assign each sample data point to the cluster whose center is nearest to it, based on the Euclidean distance.
3. Update: recalculate the mean vector (cluster center) for each cluster based on the sample data points assigned to it.
4. Iteration: repeat steps (2) and (3) until one of the following conditions is met: the predefined maximum number of iterations is reached, or the cluster centers no longer change (i.e., the mean vectors converge). Once these conditions are satisfied, the model is considered built, and the final clustering results are output.

4.2. Clustering results

Table 5 presents the characteristic ranges of the four clusters calculated by the k-means algorithm. Among these clusters, cluster 1 is characterized by low amplitude (below 90 dB) and low energy, cluster 2 features low dominant frequency (below 40 kHz) and high rise time, cluster 3 is defined by high dominant frequency (above 100 kHz), relatively high amplitude (around 90 dB), and relatively high energy, and cluster 4 exhibits high amplitude (close to 100 dB) and high energy.

4.3. Damage mode identification

Bian *et al.* (2021) classified the damage modes of fiber-reinforced concrete into three categories: matrix cracking, fiber-matrix debonding, and fiber pull-out. Based on the findings presented in Subsec. 4.2, cluster 1 is characterized by low amplitude (below 90 dB) and low energy, corresponding to matrix cracking failure. Cluster 3 is characterized by high dominant frequency (above 100 kHz), relatively high amplitude (around 90 dB), and relatively high energy, corresponding to fiber-matrix debonding failure. Cluster 4 is characterized by high amplitude (close to 100 dB) and high energy, corresponding to fiber pull-out failure. Cluster 2, by contrast, is characterized by low dominant frequency (below 40 kHz) and high rise time, corresponding to mechanical noise.

As can be seen from Table 5, the AE characteristic parameters exhibit different performance levels. Both amplitude and energy can be used to distinguish the three different damage modes, namely matrix cracking, fiber-matrix debonding, and fiber pull-out; however, relatively speaking, energy demonstrates better discrimination

Table 5. Range of clustering features (the average value is given in square brackets).

Specimen no.	AE parameter	Cluster 1	Cluster 2	Cluster 3	Cluster 4
R-0-2	Amplitude [dB]	60.1–89.6 [68.8]	69.2–100 [87.8]	70.5–100 [90.8]	99.9–100 [99.9]
	Rise time [ms]	0.01–42.5 [4.9]	0.2–104.9 [52.5]	1.1–103.8 [48.1]	0.7–104.2 [45.2]
	Count	11–5715 [4860]	1435–9182[5629]	1685–9019 [6869]	6049–8956 [7867]
	Energy [10^6 aJ]	0.001–3.6 [0.1]	0.2–80.3 [12.8]	0.3–77.1 [20]	88.5–394 [152.4]
	Dominant frequency [kHz]	26.2–125.1 [61.5]	26.9–68.4 [35.1]	59.8–123.3 [108]	27.5–109.2 [69.9]
R-0-3	Amplitude [dB]	60.1–92.4 [77.6]	72.9–100 [88.8]	76.8–100 [89.5]	92.3–100 [99]
	Rise time [ms]	0.003–69.5 [16.5]	1.1–104.8 [63.2]	0.2–104.8 [53.8]	0.6–101.3 [42.1]
	Count	21–6523 [3324]	3476–8728 [6031]	1898–9608 [6467]	4783–9706 [7795]
	Energy [10^6 aJ]	0.008–9.7 [1.3]	0.5–46.4 [9.4]	0.8–53.9 [10.2]	13.4–467 [86.4]
	Dominant frequency [kHz]	26.2–120.2 [37.9]	28.1–72 [36.2]	62.9–123.9 [106.9]	28.1–122.7 [65.2]
R-0-4	Amplitude [dB]	60.4–100 [87.6]	81.2–100 [93.2]	79–100 [90.4]	94.5–100 [99.6]
	Rise time [ms]	0.1–101.9 [30.5]	36.3–104.8 [79.2]	0.02–104.7 [53.5]	0.1–103.2 [37.5]
	Count	64–7861 [5386]	2989–8682 [6575]	4127–8717 [6875]	5454–10023 [7946]
	Energy [10^6 aJ]	0.006–32.5 [11.9]	1.8–171 [26.8]	1.1–74.8 [14]	11.6–1910 [142.3]
	Dominant frequency [kHz]	28.1–116.6 [34.5]	25.6–90.9 [34.4]	56.2–123.9 [109.9]	28.1–116.6 [65.7]
R-0-5	Amplitude [dB]	60.2–99 [73.2]	71–100 [89.9]	69.8–100 [91.4]	99.1–100 [100]
	Rise time [ms]	0.01–52 [12.4]	0.008–104.4 [52.2]	1.3–104 [52.1]	1.9–103.4 [47.8]
	Count	92–5249 [1371]	1643–9068 [5991]	1256–8968 [6803]	6109–8949 [7801]
	Energy [10^6 aJ]	0.004–18.7 [0.5]	0.2–236 [30.2]	0.2–377 [38.5]	133–2600 [522]
	Dominant frequency [kHz]	28.7–166 [56.1]	28.7–76.9 [35]	62.3–128.5 [106.7]	28.7–110.5 [52.6]
R-25-2	Amplitude [dB]	60.4–99.2 [85.2]	85.1–100 [95.4]	77.3–100 [91.1]	86.7–100 [96.7]
	Rise time [ms]	0.03–104.3 [48.4]	49.1–104.6 [79.4]	0.2–104.8 [50.3]	0.1–84.9 [25.9]
	Count	84–7954 [5562]	4128–8587 [6901]	4775–8733 [6942]	4294–8714 [6912]
	Energy [10^6 aJ]	0.005–43.6 [6.2]	2.2–218 [43.2]	1.4–113 [15.9]	4.5–1030 [66.4]
	Dominant frequency [kHz]	26.2–113.5 [35.1]	26.2–115.4 [38.1]	59.2–161.1 [110.3]	26.9–114.7 [37.1]
R-25-3	Amplitude [dB]	60.1–94.7 [74.2]	75.8–100 [89.4]	70.6–100 [88.3]	95.1–100 [99.7]
	Rise time [ms]	0.01–51.5 [10.5]	0.1–104.8 [52.1]	0.3–104.4 [54.3]	3.5–97.8 [42.6]
	Count	26–6190 [1718]	1927–9135 [6282]	2429–9091 [5954]	4474–9940 [8403]
	Energy [10^6 aJ]	0.002–7 [0.6]	0.5–109 [16.4]	0.3–78.4 [10.8]	55.1–607 [127.3]
	Dominant frequency [kHz]	27.5–121.5 [69.4]	26.9–68.4 [34.2]	69.6–125.1 [109.7]	28.1–128.1 [87.6]
R-25-4	Amplitude [dB]	62.1–97 [84.9]	80.5–100 [94.3]	80.3–100 [92.7]	91.4–100 [98.3]
	Rise time [ms]	0.05–87.7 [27.5]	36.1–104.8 [77.1]	0.3–104.2 [50.6]	0.3–100.6 [32.1]
	Count	74–7802 [5170]	4426–8673 [6617]	4952–8794 [6964]	4362–8842 [6988]
	Energy [10^6 aJ]	0.008–66.2 [5.9]	1.4–186 [30.1]	1.5–116 [16.3]	6.4–727 [107.2]
	Dominant frequency [kHz]	28.1–125.7 [47]	28.3–109.3 [37.3]	37.2–192.3 [111.5]	28.1–122.7 [38.5]
R-25-5	Amplitude [dB]	60–94.9 [72.7]	73.7–100 [88.2]	73.1–100 [89.1]	96.8–100 [99.9]
	Rise time [ms]	0.01–43.5 [8.4]	0.6–104.1 [49.8]	0.7–104.2 [51.6]	0.6–104.3 [47.6]
	Count	28–5326 [1050]	1349–9215 [5806]	1952–9253 [6322]	6074–9522 [8373]
	Energy [10^6 aJ]	0.002–9.5 [0.5]	0.3–217 [14.9]	0.3–192 [15.9]	163–968 [389]
	Dominant frequency [kHz]	29.9–123.9 [81.6]	28.1–60.4 [33.1]	75.1–123.2 [107.1]	28.1–115.4 [66.5]

performance. Taking specimen R-0-2 as an example, the average amplitude of fiber-matrix debonding is 1.32 times that of matrix cracking, and the average amplitude of fiber pull-out is 1.1 times that of fiber-matrix debonding. The average energy of fiber-matrix debonding is 200 times that of matrix cracking, and the average energy of fiber pull-out is 7.6 times that of fiber-matrix debonding. The dominant frequency can be used to distinguish mechanical noise. The average dominant frequency of mechanical noise is below 40 kHz, with most values of this cluster falling within the range of 28 dB to 80 dB, and only a few high-frequency events (above 100 dB). The count exhibits poor discrimination performance: matrix cracking failure is associated with a low count, while it is difficult to distinguish among the other three damage types using this parameter.

There is a certain correlation between compressive strength and matrix cracking. The lower the energy of matrix cracking, the higher the compressive strength of the specimen. Overall, the energy of specimens R-X-3 and R-X-5 is relatively low, while their compressive strength is relatively high.

5. Conclusions

This paper presented a study on the PCA of AE parameters and the damage modes of PPF-reinforced recycled concrete, and the following conclusions can be drawn:

1. The study showed that the Pearson correlation coefficients between the three parameters (rise time, energy, and dominant frequency) and the other parameters are mostly less than 0.6, indicating low correlation and high independence of these three parameters. The Pearson correlation coefficients among the three parameter pairs – amplitude and count, amplitude and duration, and count and duration – are mostly greater than 0.6, indicating a strong correlation between each pair. Five AE characteristic parameters, namely amplitude, rise time, count, energy, and dominant frequency, were selected for the PCA of PPF-reinforced recycled concrete.
2. Based on the SI index, the optimal number of clusters for the principal components of PPF-reinforced recycled concrete was determined to be 4. The unsupervised learning algorithm of k-means clustering was applied to identify the damage modes of PPF-reinforced recycled concrete, with four distinct damage modes being identified as follows:
   - mechanical noise, featuring low dominant frequency (below 40 kHz) and high rise time,
   - matrix cracking, characterized by low amplitude (below 90 dB) and low energy,
   - fiber-matrix debonding, exhibiting high dominant frequency (above 100 kHz), relatively high amplitude (around 90 dB) and relatively high energy,
   - fiber pull-out, characterized by high amplitude (close to 100 dB) and high energy.
3. The AE characteristic parameters exhibit varying discrimination performance. Both amplitude and energy can be used to distinguish the three distinct damage modes, namely matrix cracking, fiber-matrix debonding, and fiber pull-out; however, energy demonstrates superior discrimination performance. The dominant frequency can be used to distinguish mechanical noise, whereas the count exhibits poor discrimination performance.

It should be pointed out that some conclusions of this study need to be further verified by fiber pull-out tests.

Fundings

This study was supported by the Fujian Natural Science Foundation (Grant no. 2022J01930), the authors gratefully acknowledge this support.

Conflict of interests

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Authors' contributions

Qianxu Chen performed the analysis and contributed to data interpretation. Xin Yang conceptualized the study and wrote the original draft. All authors reviewed and approved the final manuscript.

References

1. Ashraf S., Rucka M. (2023), Microcrack monitoring and fracture evolution of polyolefin and steel fibre concrete beams using integrated acoustic emission and digital image correlation techniques, *Construction and Building Materials*, **395**: 132306, https://doi.org/10.1016/j.conbuildmat.2023.132306.

2. Ashraf S., Rucka M. (2024), Comparative study on fracture evolution in steel fibre and bar reinforced concrete beams using acoustic emission and digital image correlation techniques, *Case Studies in Construction Materials*, **20**: e03359, https://doi.org/10.1016/j.cscm.2024.e03359.

3. Bian C., Wang J.Y., Guo J.Y. (2021), Damage mechanism of ultra-high performance fibre reinforced concrete at different stages of direct tensile test based on acoustic emission analysis, *Construction and Building Materials*, **267**: 120927, https://doi.org/10.1016/j.conbuildmat.2020.120927.

4. Chkhachirou M., El-Hassan H., El-Maaddawy T. (2025), Durability of BFRP bars embedded in geopolymer concrete under hygrothermal exposure and sustained loading, *Case Studies in Construction Materials*, **23**: e05571, https://doi.org/10.1016/j.cscm.2025.e05571.

5. Džolan A., Fischer O., Niedermeier R. (2024), Analyses of the fatigue behavior of carbon short-fiber-reinforced concrete (CSFRC) under tension and flexion, *Construction and Building Materials*, **453**: 139058, https://doi.org/10.1016/j.conbuildmat.2024.139058.

6. Ingle V.V., Prem P.R. (2025), Acoustic emission examination of 3D printed ultra-high performance concrete with and without coarse aggregate under fresh and hardened states, *Journal of Building Engineering*, **111**: 113491, https://doi.org/10.1016/j.jobe.2025.113491.

7. Kantekin Y., Bakir B.B. (2025), Joint shear model for fiber reinforced concrete beam-column connections, *Journal of Building Engineering*, **101**: 111833, https://doi.org/10.1016/j.jobe.2025.111833.

8. Kouta N., Saliba J., Saiyouri N. (2021), Fracture behavior of flax fibers reinforced earth concrete, Engineering Fracture Mechanics, **241**: 107378, https://doi.org/10.1016/j.engfracmech.2020.107378.

9. Ministry of Housing and Urban-Rural Development of the PRC (2019), *Standard for test methods of concrete physical and mechanical properties* (Standard No. GB/T 50081-2019).

10. Sagar R.V., Samadhan S.A., Kundu T. (2025), Tensile stress-crack width relationship for steel fiber reinforced concrete under mode I fracture, *Mechanics Research Communications*, **144**: 104378, https://doi.org/10.1016/j.mechrescom.2025.104378.

11. Saha I., Sagar R.V. (2021), Classification of the acoustic emissions generated during the tensile fracture process in steel fibre reinforced concrete using a waveform-based clustering method, *Construction and Building Materials*, **294**: 123541, https://doi.org/10.1016/j.conbuildmat.2021.123541.

12. Tayfur S., Alver N., Adbi S., Saatci S., Ghiami A. (2018), Characterization of concrete matrix/steel fiber debonding in an SFRC beam: Principal component analysis and k-mean algorithm for clustering AE data, *Engineering Fracture Mechanics*, **194**: 73–85, https://doi.org/10.1016/j.engfracmech.2018.03.007.

13. Umar H.A., Zeng X.H., Long G.C., Tang Z., Lan X.L., Zhu H.S. (2023), Synergistic effects of asphalt emulsion and fiber reinforcement on fracture properties and energy absorption of self-compacting concrete, *Theoretical and Applied Fracture Mechanics*, **127**: 104043, https://doi.org/10.1016/j.tafmec.2023.104043.

14. Wang C.Q., Zhang Y.Y., Ma Z.M., Wang D.J. (2022), Hysteretic deteriorating behaviors of fiber-reinforced recycled aggregate concrete composites subjected to cyclic compressive loadings, *Journal of Building Engineering*, **49**: 104087, https://doi.org/10.1016/j.jobe.2022.104087.

15. Wen Y.J. (2025), *Research on damage pattern identification and evolution mechanism of steel-concrete structures based on acoustic emission and machine learning* [in Chinese], M.Sc. Thesis, Guangxi University.

16. Zaki Y.A., Abouhussien A.A., Hassan A.A.A., Ismail M.K., AbdelAleem B.H. (2023), Crack detection and classification of repaired concrete beams by acoustic emission monitoring, *Ultrasonics*, **134**(5): 107068, https://doi.org/10.1016/j.ultras.2023.107068.

17. Zheng Q.M., Li C., He B., Jiang Z.W. (2022), Revealing the effect of silica fume on the flexural behavior of ultra-high-performance fiber-reinforced concrete by acoustic emission technique, *Cement and Concrete Composites*, **131**(6): 104563, https://doi.org/10.1016/j.cemconcomp.2022.104563.

Archives of Acoustics Vol. 51, No. 2, pp. 293–299 (2026), https://doi.org/10.24423/archacoust.2026.4396

Research Paper

The Mechanism of Formation of the Cutoff Frequency in an Acousto-Optic Delay Line and Some Proposals for its Measurement

Afig Hasanov, Ruslan Hasanov, Elgun Aghayev*, Rovshan Ahmadov

Department of Radio Electronics, National Aviation Academy
Baku, Azerbaijan

*Corresponding Author: eaghayev@naa.edu.az

Received December 14, 2025; revised February 23, 2026; accepted March 20, 2026; available online April 1, 2026; version of record June 3, 2026; published issue June 24, 2026.

The structure of the acousto-optic delay line and possible areas of its application are discussed. The necessity of determining the cutoff frequency of the acousto-optic delay line in all areas of its application is substantiated. The relationship between the cutoff frequency of an electric circuit and its time constant is discussed. The obtained result is extrapolated to the acousto-optic delay line, in which the cutoff frequency is formed by a completely different mechanism. The mechanism of cutoff frequency formation in the acousto-optic delay line is discussed. It is shown that the cutoff frequency in the acousto-optic delay line is formed due to the finite velocity of interaction of the acoustic wave with the light beam in the photoelastic medium. Two methods for measuring the cutoff frequency of the acousto-optic delay line are discussed: the method of system-parametric measurement and the method of cursor measurement. The system-parametric method for measuring the cutoff frequency of the acousto-optic delay line is implemented based on the known values of the laser beam diameter and the velocity of propagation of the acoustic wave in the photoelastic medium. The cursor method for measuring the cutoff frequency of the acousto-optic delay line is implemented based on the parameters of the oscillogram of its response to the input action in the form of a rectangular pulse. Theoretical and experimental aspects of the application of these methods are discussed. Corresponding numerical examples and experimental results are given.

Keywords: acousto-optic, delay line, cutoff frequency, transient response, system-parametric measurement, Bragg diffraction.

1. Introduction

An acousto-optic delay line (AODL) is implemented based on the photoelastic effect and is used to process signals in the time domain. AODL differs from other types of acousto-optic processors in that information it contains is extracted by a light beam whose diameter is significantly smaller than the size of the optical aperture of the interaction medium. In devices of this class, efficient signal processing in the time domain is due to the low velocity of propagation of the acoustic wave in the photoelastic medium (PEM), the ability to regulate the delay time of the electrical signal by simple mechanical movement of the interaction medium, and the ability to synthesize PEMs of sufficiently large sizes. The AODL can be used in radio engineering systems for various purposes, in radar target simulators (Diewald *et al.*, 2018; Okoń-Fąfara *et al.*, 2019), and also for solving other radar problems associated with the formation of time intervals (da Silva *et al.*, 2024; Daggula, Bevara, 2024; Russell *et al.*, 2025; Wu, 2025). Diewald *et al.* (2018) described a system capable of simulating the behavior of radar targets. Okoń-Fąfara *et al.* (2019) discussed the development of a radar air situation map simulator that can be used to train military radar operators, to evaluate the functionality of new radar

systems and signal processing algorithms without the need for physical equipment, etc. In both cases, digital devices capable of generating and processing signals with a limited spectrum are used. However, generating and processing broadband signals is very difficult or impossible.

The use of AODL in solving the problems considered in the indicated works allows for a significant expansion of the technical capabilities of the corresponding devices and systems. In all cases of application of the AODL, one of the main parameters is its cutoff frequency. The cutoff frequency of an arbitrary electrical circuit is defined as the frequency at which the output signal power is reduced by half from the original value. This parameter is determined by the structure of the electrical circuit and the reactive parameters of each node in its composition. Any electric circuit responds to an input action not instantly, but with a certain delay. This delay is usually estimated by the time constant of the transient response of the electric circuit. The time constant of the transient response of the electric circuit τ is defined as the time interval from the beginning of the formation of the output response to the input action in the form of a single-step to the moment when the value of this response reaches the level of $(1-1/e)\cdot 100 = 63.2\,\%$ of the steady-state value. The cutoff frequency of the electric circuit f_c is determined by the known value of its time constant τ as follows:

$$f_c = 1/(2\pi\tau). \tag{1}$$

From Eq. (1) it follows that, using any method that allows one to determine the time constant of an electric circuit, one can calculate its cutoff frequency. For example, in the case of an RC circuit, the time constant is defined as the product of the nominal values of the elements R and C. Accordingly, the formula for calculating the cutoff frequency (Eq. (1)) takes the following form: $f_c = 1/2\pi RC$. A similar statement is true for all electric (including radio engineering) circuits whose equivalent circuits can be represented as an RC circuit.

The cutoff frequency is also formed in the AODL, despite the obvious absence of reactive elements in its structure. The article is devoted to the structural interpretation of the mechanism for forming the cutoff frequency of an acousto-optic delay line and solving the problem of its measurement.

2. Material and method

In the AODL, the processed analog signal $u_{\text{in}}(t)$ is fed to the first input of the amplitude modulator (AM) (Fig. 1). The carrier oscillation from the high-frequency generator (HFG) is fed to the second input of the AM. An amplitude-modulated radio-frequency signal is formed at the AM output. It should be noted that, depending on the nature of the problem being solved, balanced amplitude modulation, frequency modulation, or amplitude manipulation can also be used. The amplitude-modulated signal at the AM output is described by the expression:

$$u_{\text{AM}}(t) = U_0[1 + m\cdot s(t)]\cdot\cos(\omega_0 t),$$

where U_0 is the amplitude, ω_0 is the carrier oscillation frequency, $s(t)$ is the modulating process, which changes within the limits $s(t) = \pm 1$. In our case $s(t) = u_{\text{in}}(t)/[u_{\text{in}}(t)]_{\max}$, where $[u_{\text{in}}(t)]_{\max}$ is the maximum value of the processed analog signal $u_{\text{in}}(t)$.

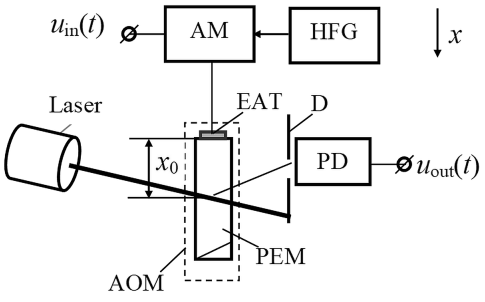

Fig. 1. Schematic diagram of an acousto-optic delay line.

The amplitude-modulated signal is transmitted to an electro-acoustic transducer (EAT) attached to the end of the PEM. The EAT excites an acoustic wave in the PEM, which propagates with a velocity v that is ~10^5 times smaller than the propagation velocity of an electromagnetic wave. A cell consisting of the PEM and the EAT attached to its end is called an acousto-optic modulator (AOM) (Davis, 2014; Molchanov *et al.* 2015). In this case, the frequency of the carrier oscillation ω_0 generated in the HFG is selected in the range of operating frequencies of the AOM, which, as a rule, is 40 % to 60 % of its central frequency (Yushkov *et al.*, 2024). The central frequency of the AOM can be selected from tens of MHz to units of GHz. It follows that when constructing an AODL, the value of the carrier oscillation frequency does not impose any special restrictions on its operational and technical characteristics, and it is possible to use existing ready-made units, in particular, the AOM. In this case, AM and HFG are used only to transfer the spectrum of the processed analog signal to the operating frequency range of the AOM.

The requirement for the maximum frequency of the modulating process $s(t)$ is formed based on the known value of the AODL cutoff frequency. The AODL cutoff frequency must be equal to or greater than the maximum frequency of the processed analog signal $u_{\text{in}}(t)$. In other words, the design and technological characteristics of the AODL must be adapted to the maximum frequency of the processed analog signal $u_{\text{in}}(t)$.

The circuit in Fig. 1 uses the Bragg diffraction mode. A laser beam of diameter d passing through the AOM is modulated by diffraction on inhomogeneities in permittivity caused by deformations of the PEM material under the action of an acoustic wave. Part of the light beam is deflected into the diffraction order. The deflected light falls through the hole in the diaphragm (D) onto the surface of the photodetector (PD) and is detected. As a result, a voltage $u_{\text{out}}(t)$ is formed at the PD output, which repeats the shape of the processed input signal $u_{\text{in}}(t)$ and lags behind it in time:

$$t_d = x_0/v, \tag{2}$$

where x_0 is the distance from the EAT to the acousto-optic interaction point in the PEM.

In other words, ideally for AODL the equality:

$$u_{\text{out}}(t) = c \cdot u_{\text{in}}(t - t_d), \tag{3}$$

holds, where c is a constant factor.

In a real AODL, equality (Eq. (3)) is fulfilled with distortions acceptable for practice in the frequency band limited by the cutoff frequency. This parameter is formed due to the finite speed of physical processes in the nodes that participate in the formation of the response at the AODL output. Abstracting from secondary factors, it can be assumed without proof that in the chain of nodes participating in the formation of the AODL cutoff frequency, the AOM and PD have an incomparably greater influence. When using a high-speed photomultiplier, for example a photomultiplier based on the FEU-114 type, its influence on the cutoff frequency of the AODL can also be neglected. Therefore, based on this interpretation of the operation of an AODL, it can be assumed that, under certain conditions, its cutoff frequency is formed by the final speed of intersection of a light beam with an acoustic wave in a photoelastic medium (Muromets *et al.*, 2016; Hasanov *et al.*, 2022). With these interacting wave parameters (the diameter of the light beam d and the velocity of propagation of the acoustic wave in the photoelastic medium v), the following expression can be obtained for the transient response of the AODL (Hasanov *et al.*, 2019):

$$g(t) = \frac{8}{\pi\left(\frac{d}{v}\right)^2} \cdot \int_{t_d}^{t} \sqrt{\frac{d}{v}(\xi - t_d) - (\xi - t_d)^2}\, \mathrm{d}\xi, \qquad \text{for} \quad t_d \le t \le t_d + \frac{d}{v}. \tag{4}$$

Equation (4) provides a concrete mathematical description of how an AODL responds to a single-step input signal, and allows us to predict and calculate the response of an AODL to virtually any conceivable input signal using the Duhamel integral. This makes Eq. (4) and the Duhamel integral a powerful combination for analyzing and predicting the behavior of an AODL under various operating conditions.

The graph of the transient response of the AODL, constructed using Eq. (4), with the parameter values $t_d = 0.1$ μs, $v = 3.63$ km/s, $d = 2.2$ mm is shown in Fig. 2.

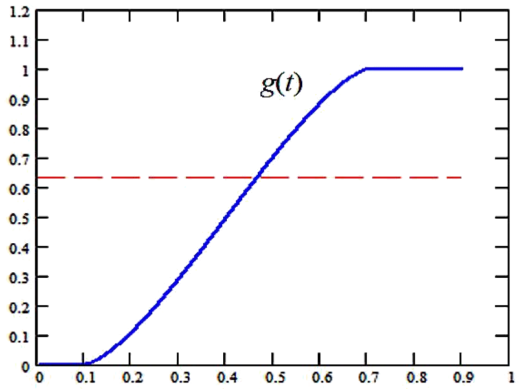

FIG. 2. Graph of the transient response of the AODL.

The time constant of the AODL (and, accordingly, its cutoff frequency) can be calculated both by the formula of the transient response (Eq. (4)) and by its graph (Fig. 2), which is the basis of two methods for measuring the cutoff frequency of the AODL. These methods are discussed further.

3. System-parametric method for measuring the cutoff frequency of an AODL

The system-parametric method for measuring the AODL cutoff frequency consists of determining the AODL time constant by its transient response based on known values of the parameters of the interacting beams (acoustic and optical). Then, the AODL cutoff frequency is calculated based on the known value of the time constant. From Eq. (4), the time constant of the AODL is determined as follows:

$$\tau = t_1 - t_d, \tag{5}$$

where $t_1 = t_d + \tau$ is the moment in time when the right side of Eq. (4) is equal to 0.632, i.e., the equality $g(t_1) = 0.632$ is satisfied.

The cutoff frequency f_c of a specific type of AODL does not depend on the value of the delay time t_d, which is determined by Eq. (2). From a joint analysis of Eq. (1) to Eq. (5), it is easy to conclude that, using known values of the parameters v and d, it is possible to calculate the cutoff frequency of the AODL with sufficient accuracy for engineering calculations. Calculations are easily performed in the Mathcad environment. Calculations are performed for the following values of the design parameters: $t_d = 0.1\,\mu s$, $v = 3.63\,km/s$, $d = 2.2\,mm$. Equating the right side of Eq. (4) to 0.632, we find the time constant of the AODL τ. Thus, for the time constant of the AODL we obtain the following value: $\tau = 0.366\,\mu s$. Accordingly, the cutoff frequency of the AODL, calculated using Eq. (1), is

$$f_c = 1/(2\pi \cdot 0.366) = 0.435\,MHz.$$

4. Cursor method for measuring the cutoff frequency of an AODL

The method of cursor measurement of the AODL cutoff frequency consists of determining its time constant based on the parameters of the oscillogram of its response to the pulse input signal. Then, based on the known value of the time constant, the AODL cutoff frequency is calculated.

In the experimental research model, a rectangular pulse with the required parameters is formed in the G5-54 pulse generator. The pulse from the G5-54 generator output modulates the oscillations of the G4-107 high-frequency generator (operates in the external pulse modulation mode) and synchronizes the MSO4052 oscilloscope. The oscillation frequency of the G4-107 generator is selected equal to the central frequency of the AOM, which in our experiments is 80 Hz. The deflected light in the rear focal plane of the AOM is recorded

by a PD based on a FEU-114 photoelectron multiplier. The devices in the laboratory setup are connected to each other by radio frequency cables.

The oscillograms of the voltages at the input and output of the AODL with the parameters t_d = 0.1 μs, v = 3.63 km/s, d = 2.2 mm, $\tau_i \approx 1.5$ μs are shown in Fig. 3.

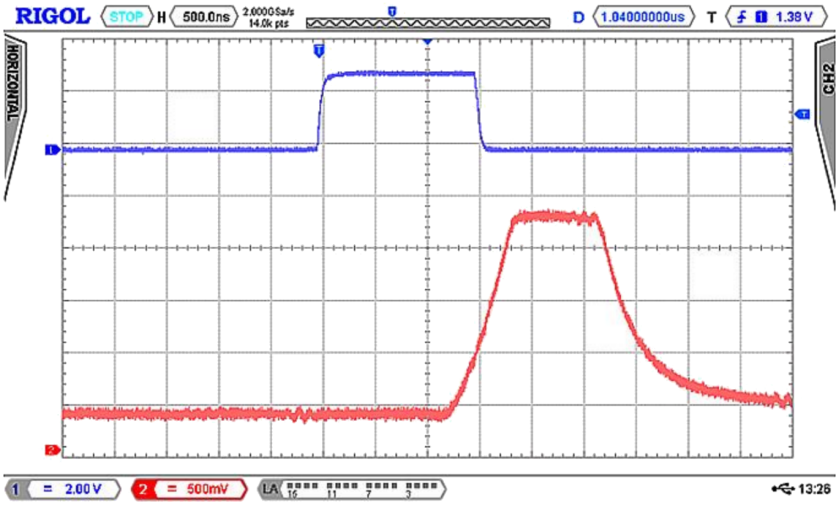

Fig. 3. Oscillograms of pulses at the input (1) and output (2) of an acousto-optic delay line.

The duration of the input pulse (determined from the oscillogram at a level of 0.5 of the maximum value) is equal to $\tau_i \approx 1.5$ μs. The time constant of the transient response, i.e., the time during which the response – the output voltage changes from 0 % to 63.2 % of its maximum value (≈1.9 V), is equal to approximately 380 ns, which coincides with the calculated values of the time constant of the transient response. The experimentally measured value of the AODL cutoff frequency is $1/(2\pi \cdot 0.38)$ = 419 kHz. Thus, the spread from the average value in the obtained results of system-parametric (435 kHz) and cursor (419 kHz) measurements does not exceed 5 %.

The discrepancy between the results of system-parametric and cursor measurements is primarily due to the influence of the photodetector. In the case of system-parametric measurements, the influence of the photodetector is not taken into account. However, the photodetector does contribute to the results of cursor measurements.

5. Results and discussion

When constructing an AODL for processing an analog signal, it is important to optimize the cutoff frequency in the context of the maximum signal frequency. The problem can be solved directly or indirectly. Direct measurement of the AODL cutoff frequency can be performed using an experimentally measured amplitude-frequency characteristic. To measure the amplitude-frequency characteristic, appropriate devices are needed. In addition, this takes a lot of time, especially if several measurements are required. Therefore, indirect measurement is more preferable. Indirect measurement of the cutoff frequency is possible using the system-parametric and cursor methods. The system-parametric method for measuring the AODL cutoff frequency appeals to a priori known values of the laser beam diameter and the acoustic wave propagation velocity in the PEM. The cursor method of measuring the AODL cutoff frequency is implemented by the parameters of its response to the input action in the form of a rectangular pulse. Accordingly, these methods can be used at various stages of the AODL development. At the design stage, it is preferable to use the system-parametric measurement, at the testing stage – the cursor measurement. From the interpretation of the measurements, it follows that the results obtained by different methods differ by no more than four percent. These errors are mainly due to the fact that the measurements were made using an oscillogram. With direct measurements, the accuracy will be much higher, since the digital oscilloscope provides measurements of time parameters with an accuracy of up to 0.0001 %.

In the presented methods for measuring the AODL cutoff frequency, the influence of the photodetector is not discussed. This is due to the fact that a high-speed photodetector is assumed to be used a priori. In this

case, a photodetector of the FEU-114 type was used, for which the rise time of the transient response, according to the passport data, is ≤9 ns. Subtracting this value from the cursor measurement result (380 ns) yields 371 ns. The corresponding AODL cutoff frequency is $1/(2\pi \cdot 0.371) = 429$ kHz, which is slightly different from the system-parametric measurement result (435 kHz). All this confirms the postulate that the AODL cutoff frequency is determined by the interaction time of the light and elastic wave.

6. Conclusion

One of the main parameters of the AODL is its cutoff frequency. This parameter is taken into account at the design stage, during operation and when analyzing the output results. The methods of system-parametric measurement and cursor measurement of the AODL cutoff frequency allow solving this problem quite simply and correctly. Based on these methods, it is possible to select the type of PEM material and the parameters of the light source (laser). The method of system-parametric measurement is used at the design stage. At the same time, this method can also be used to test the results of experimental studies.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions

Afig Hasanov developed the method, prepared figures, and wrote the manuscript. Ruslan Hasanov, Elgun Aghayev, Rovshan Ahmadov conducted measurements and provided technical approval. All authors reviewed and approved the final manuscript.

References

1. da Silva R.E., Manuylovich E., Sahoo N., Franco M.A.R., Bartelt H., Webb D.J. (2024), All-fiber fast acousto-optic temporal control of tunable optical pulses, *Optical Fiber Technology*, **87**: 103877, https://doi.org/10.1016/j.yofte.2024.103877.
2. Daggula R., Bevara V. (2024), An ultra-low power QCA based vedic multiplier for digital radar application, *e-Prime – Advances in Electrical Engineering, Electronics and Energy*, **9**: 100695, https://doi.org/10.1016/j.prime.2024.100695.
3. Davis C.C. (2014), *Lasers and Electro-Optics*, Cambridge University Press.
4. Diewald A.R., Steins M., Müller S. (2018), Radar target simulator with complex-valued delay line modeling based on standard radar components, *Advances in Radio Science*, **16**: 203–213, https://doi.org/10.5194/ars-16-203-2018.
5. Hasanov A. *et al.* (2022), Development of an axonometric model of photoelastic interaction in an acousto-optic delay line and its approbation, *Technology Audit and Production Reserves*, **5**(2(67)): 38–45, https://doi.org/10.15587/2706-5448.2022.267782.
6. Hasanov A.R., Hasanov R.A., Ahmadov R.A., Agayev E.A. (2019), Time- and frequency-domain characteristics of direct-detection acousto-optic delay lines, *Measurement Techniques*, **62**: 817–824, https://doi.org/10.1007/s11018-019-01700-3.
7. Molchanov V.Ya. *et al.* (2015), *Theory and Practice of Modern Acousto-Optics* [in Russian], Publishing House MISiS.

8. Muromets A.V., Voloshinov V.B., Kononin I.A. (2016), Transmission characteristics of acousto-optic filter using sectioned transducer, *Applied Acoustics*, **112**: 221–225, https://doi.org/10.1016/j.apacoust.2016.04.008.

9. Okoń-Fąfara M., Kawalec A.M., Witczak A. (2019), Radar air picture simulator for military radars, [in:] *XII Conference on Reconnaissance and Electronic Warfare Systems*, **1105519**, https://doi.org/10.1117/12.2525032.

10. Russell R.S., Anderson B.E., Denison M.H. (2025), Using time reversal with long duration broadband noise signals to achieve high amplitude and a desired spectrum at a target location, *Applied Acoustics*, **236**: 110744, https://doi.org/10.1016/j.apacoust.2025.110744.

11. Wu M., Ma J., Zhang Q. (2025), Photonic-assisted angle-of-arrival measurement system for both broadband and single-frequency radar signals, *Optics Communications*, **577**: 131392, https://doi.org/10.1016/j.optcom.2024.131392.

12. Yushkov K.B., Naumenko N.F., Molchanov V.Ya. (2024), Design of a broadband acousto-optic filter using bulk acoustic wave beam steering with an interdigital transducer, *Results in Physics*, **59**: 107575, https://doi.org/10.1016/j.rinp.2024.107575.

Archives of Acoustics Vol. 51, No. 2, pp. 301–316 (2026), https://doi.org/10.24423/archacoust.2026.4038

Research Paper

Shaping the Soundscape: Exploring the Influence of Building Layout on Outdoor Acoustic Environments

Sami Hamouta[(1)*], Atef Ahriz[(2)], Noureddine Zemmouri[(3)], Ahmed Mansouri[(4)]

[(1)] *Laboratory of Civil Engineering and Hydraulics (LGCH)*
Department of Architecture, 8 Mai 1945 Guelma University
Guelma, Algeria

[(2)] *Applied Civil Engineering Laboratory (LGCA)*
Department of Architecture, University of Tebessa
Tebessa, Algeria

[(3)] *Laboratory of Design and Modeling of Architectural and Urban Forms and Ambiances (LACOMOFA)*
Department of Architecture, University of Biskra
Biskra, Algeria

[(4)] *Department of Architecture, University of Batna 1*
Batna, Algeria

*Corresponding Author: hamouta.sami@univ-guelma.dz

Received April 18, 2024; revised August 23, 2025; accepted January 12, 2026;
available online January 19, 2026; version of record April 16, 2026; published issue June 24, 2026.

This study investigated the influence of building layout on the outdoor acoustic environment through field measurements conducted in four courtyards at the University of Batna 1. Acoustic parameters including sound pressure level (SPL) attenuation, reverberation time (RT), early decay time (EDT), clarity (D50), and the rapid speech transmission index (RaSTI) were evaluated. Results showed that square-shaped courtyards retained sound the longest (RT20 exceeding 2.3 s at 1000 Hz), U-shaped courtyards exhibited the most irregular reverberation patterns, and linear courtyards provided the most stable sound decay. The D50 and RaSTI values were highest in linear courtyards, indicating superior speech intelligibility, while square and U-shaped layouts demonstrated reduced intelligibility due to extended reverberation. The SPL attenuation was also more consistent in linear configurations compared to the variable patterns observed in enclosed geometries. These findings demonstrate that building form plays a decisive role in shaping outdoor acoustic conditions and highlight the importance of considering acoustic performance in early design decisions. The results are broadly applicable to the planning of courtyards, plazas, and semi-enclosed urban spaces. Future work should explore additional variables such as building height, façade materials, vegetation, and seasonal effects to develop comprehensive guidelines for acoustically optimized outdoor environments.

Keywords: outdoor acoustic environment, architectural design optimization, building geometry and acoustics, noise mitigation strategies, acoustic comfort in urban spaces.

1. Introduction

In urban outdoor environment, the shifting toward the dependence on more mechanization has led to a gradual acceptance of noise as a fact (Wang *et al.*, 2005). However, noise may have both immediate and long-term detrimental impacts on human health and the environment, with real and perceived repercussions, inter-

fering with sleep, concentration, communication, and recreation (World Health Organization, 2018; Goines, Hagler, 2007). The university environment, for instance, encompasses an urban complexity relevant to this dilemma.

In the last decade, sound environment at the cognitive performance contexts has received considerable attentions (Çolakkadioğlu *et al.*, 2018; Goswami *et al.*, 2011; Su *et al.*, 2013; Xie *et al.*, 2011). These contexts have become more and more exposed to high level of environmental noise (Zannin *et al.*, 2013; Zannin, Ferraz, 2016; Zannin, Zwirtes, 2009). The auditory environment has an impact on students' behavior and comprehension, i.e., loud environments are not conducive to learning, making instruction laborious, and generating annoyance and difficulty focusing (Çolakkadioğlu *et al.*, 2018; Su *et al.*, 2013; Zannin, Ferraz, 2016).

University outdoor spaces are essentially designed to be aiming at preserving the restorative and relaxation of students (Gulwadi *et al.*, 2019). Nevertheless, outdoor areas enclosed by buildings are the first ones to be exposed and impacted by noise sources. Buildings provide several complex acoustic effects when sound travels through the air, affecting both the transient sound levels associated with reverberation time (RT) and the continuous noise levels, such as sound pressure level (SPL), often generated by road traffic (Yang *et al.*, 2013). The occurrence of many reflections, diffractions, and diffusions is contingent upon factors such as dimensions, irregularity, material characteristics, architectural arrangement, and ground surfaces of the structure. They can affect auditory comfort for both leisure and relaxation in open-air environments. They also can impact indoor facilities, such as classrooms, libraries, and laboratories that might be exposed to high noise background of those outdoor spaces. Hence, creating outdoor places that have pleasant acoustics may enhance the overall quality of life for educational, instructional, and relaxation purposes.

The shape of the built environment greatly influences the acoustic properties of outdoor areas (Benameur *et al.*, 2022; Bouzir, Zemmouri, 2017; Montalvão Guedes *et al.*, 2011; Oliveira, Silva, 2011; Silva *et al.*, 2014; Wang, Kang, 2011). Its various features have the potential to modify noise levels, building layout shape and arrangement, being one of these key elements, can contribute to altering the acoustic parameters of the sound environment.

Previous research endeavors have delved into the impact of fabric environment features related to the building layout that shape the acoustic ambiance of outdoor spaces. Ariza-Villaverde *et al.* (2014), Lee, Kang (2015), Thomas *et al.* (2013) investigated the effect of street width and building height. The findings indicate that the H/W ratio had an impact on the variance of sound characteristics. Echevarria Sanchez *et al.* (2016) conducted a study on the influence of building shape on the street canyon effect and noise exposure, the results indicated that flat vertical, flat upwardly inclined, flat downwardly inclined, upwardly stepped convex, downwardly stepped, and concave may significantly influence individuals' noise exposure. The study conducted by Eggenschwiler *et al.* (2022) examined the impact of building rotation, specifically the orientation of walls (parallel vs. nonparallel), on the perception of noise discomfort. Rotating the building (which leads to walls that are not parallel) was shown to be linked to reduced noise nuisance compared to the initial orientation with parallel walls. Although the decrease in sound intensity contributed to this outcome, the beneficial impact also persisted when the sound level was the same for both rotating and parallel structures.

Additional research has also shown that different morphological elements of building layouts might impact the acoustic environment. The studies by Flores *et al.* (2017) and Yang *et al.* (2013; 2017) focused on investigating the effect of configuration and disposition of building on acoustical parameters such as RT, early decay time (EDT), definition (D50), and rapid speech transmission index (RaSTI), as well as the attenuation of the SPL in outdoor spaces. They highlight that the configuration and disposition of the building such as linear-shaped, square-shaped, U-shaped, and parallel-shaped has a crucial effect on sound environment.

Han *et al.* (2018) aimed to examine the impact of geographical landscape features on Urban Environment Noise (UEN) and traffic noise in the Shenzhen metropolitan area of China. The study revealed substantial correlations between urban morphology and regional traffic noise levels. The design and structure of buildings have a substantial correlation with regional noise (RN). The arrangement of buildings is associated with traffic noise (TN), and continuous and interconnected structures along the sides of highways are more efficient in reducing the impacts of TN. The scattered distribution and uneven forms of buildings aid in the reduction of RN.

Buildings are more efficient in mitigating noise when they are dispersed across metropolitan areas, rather than being concentrated in one area.

Prior investigations, using site measurement, have been carried out to assess the impact of buildings on RT and noise levels in outdoor areas (Aylor *et al.*, 1973; Flores *et al.*, 2017; Picaut *et al.*, 2005; Steenackers *et al.*, 1978; Thomas *et al.*, 2013; Wiener *et al.*, 1965; Yang *et al.*, 2013; 2017; Yeow, 1977; Zuccherini Martello *et al.*, 2015). In conclusion, the site measurements' findings show that buildings in urban areas are a contributing factor to rising noise levels and RT because of multiple reflections. Street width and the acoustic characteristics of ground surfaces, building layouts, and façades are some of the factors that affect these reflections.

While outdoor spaces at universities may have certain similarities with urban streets, squares, and built-up regions, they exhibit diverse layouts, sorts, and sizes. Additional investigation is necessary to examine the acoustic properties of outdoor areas inside these specific cognitive structures, since there are variations in the arrangement of buildings, materials used, and façade layout.

The objective of this research is to analyze the acoustic properties of outdoor spaces surrounding by buildings by examining data collected from four outdoor areas at University of Batna 1, each with distinct building layouts. The outdoor areas were classified into four distinct building layouts: U-shaped, square-shaped, linear-shaped, and L-shaped. The RT, EDT, and SPL attenuation were assessed based on the on-site measurements, taking into account the distances between the sound source and the receiver. An analysis was conducted on the features of room acoustical factors utilizing the RaSTI and D50, both of which are associated with speech intelligibility.

2. Methods

2.1. Description of the case study

This study aims to investigate the impact of building façades on the sound environment in four outdoor spaces within the campus of the University of Batna 1, situated in Batna (Aurès region), North East of Algeria.

The outdoor areas were chosen for their close proximity to university buildings and their regular use by students, as seen in Fig. 1 and Fig. 2. Furthermore, the selection process took into account the building components

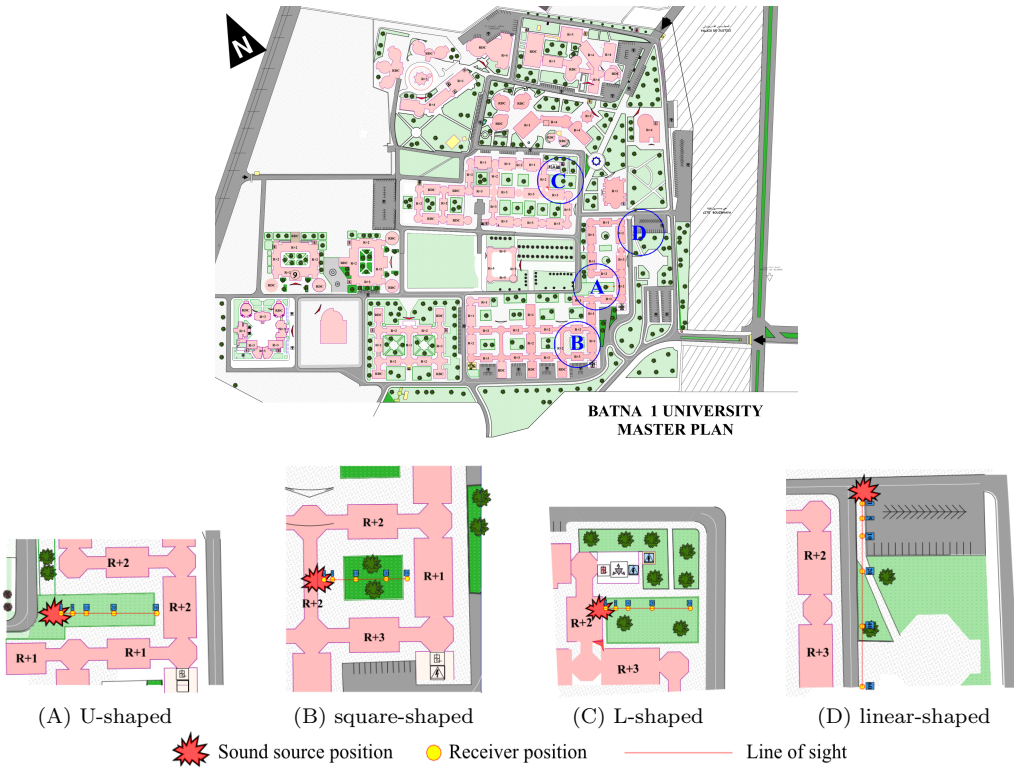

Fig. 1. University of Batna 1, master plan and measurements' station's location.

(A) U-shaped (B) square-shaped

(C) L-shaped (D) linear-shaped

FIG. 2. Photographs of each measurement stations.

and layout forms. These spaces share a set of architectural features. Among these features an almost square shape measuring about 40 m by 40 m, façades composed of concrete walls and large glass windows. Façade heights range from 2 to 3 floors. However, each outdoor space has a different building layout. These architectural aspects of the structure may be seen as surfaces that reflect sound, resulting in a longer RT and an increase in SPL owing to intense reflections compared to an open area. Building layouts surrounding outdoor spaces are categorized into four types: square-shaped (i.e., □), U-shaped (i.e., U), L-shaped (i.e., L), and linear-shaped (i.e., –).

2.2. Measurement protocol

Figure 3 displays the workflow of this research to analyze the impulse response, including RT, EDT, D50, RaSTI, and SPL.

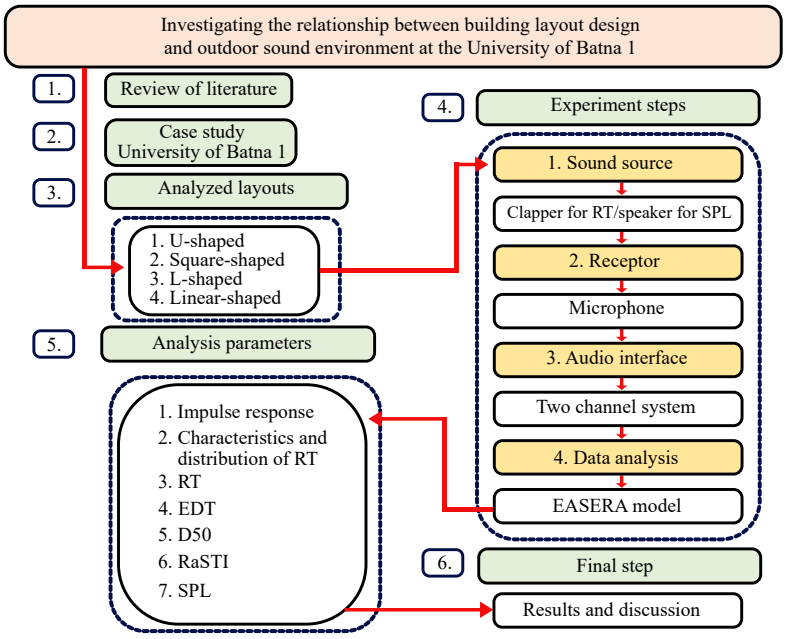

FIG. 3. Investigation workflow.

The measurement of SPL attenuation with distance was conducted using the Scarlet Solo Focusrite audio interface, a 1/2 inch measuring microphone (Dayton audio type EMM-6), and a real-time analyzer (RTA) inside the ESERA software developed by AFMG. The receiver was calibrated before the measurement and set up to 1.5 m of height.

The measurement used white noise as the sound source, which was produced by a directional speaker at a height of 1.5 m. The signal-to-noise ratio (SNR) for the measurement was 53 dB at a distance of 1 m from the source. This indicates that there was enough sound power to accurately measure the SPL attenuation at distances up to 40 m between the source and the receiver. The SNR measured at a distance of 40 m from the source was around 25 dB.

The impulsive signal was generated using a starting clapper, selected because it provides higher sound levels relative to background noise compared to an omni-directional speaker. The clapper produces a broadband impulsive signal with a nearly omnidirectional radiation pattern in the horizontal plane, making it well suited for outdoor impulse response measurements. The impulse responses for the starting clapper were captured via the Scarlet Solo Focusrite audio interface and a 1/2 inch measurement microphone (Dayton audio type EMM-6). The acoustical characteristics, such as RT, EDT, clarity (D50), and speech intelligibility (RaSTI), were examined using the EAZERA software from AFMG. This program has a noise compensation algorithm that minimizes the impact of background noise on the computation of RT. Both the source and the receiver were positioned at a height of 1.5 units above the ground. Each measurement represented the average value obtained from five consecutive claps performed in succession. Prior to each measurement, the microphone underwent calibration.

Before and after each measurement session, the microphones and acquisition system were calibrated using acoustic calibrator (94 dB at 1 kHz). This ensured the accuracy and reliability of the recorded acoustic parameters across all measurement points.

All measurements, conducted on the same winter day under the meteorological conditions detailed in Table 1, ensured consistent conditions and minimized the influence of weather variations between sites, thereby guaranteeing comparability across the different measurement zones.

Table 1. Weather conditions of each measurement zone.

Weather condition	U-shaped	Square-shaped	Linear-shaped	L-shaped
Temperature [°C]	12.1	12.7	13.1	12.7
Humidity [%]	35	35	48	35
Wind speed [m/s]	<2.2	<5.25	<5.59	<5.25

Meteorological data were verified to ensure compliance with guidance and regulations for outdoor acoustic measurements. The slight exceedances in two cases (5.25 m/s and 5.59 m/s) are considered negligible in terms of their potential influence on the results.

Ground surface materials play a significant role in outdoor sound propagation and reflections, as they influence both absorption and scattering. In this study, however, all investigated spaces had ground surfaces composed mainly of natural soil and green cover. Since this condition was uniform across the sites, its effect on differentiating the acoustic outcomes is minimal, ensuring that the observed variations are more strongly related to the architectural configuration.

The number and position of the source and recipient points in each measurement zone are outlined in Table 2. Figure 1 depicts the positions of source to receiver points in the four regions, with a total of 21 sites used for

Table 2. Number and position of the source and receiver points at every measurement areas.

Type of the building layouts	Number of sources	Number of receivers	Source-receiver distance [m]	Measurement parameter	
				SPL attenuation	Impulse response
U-shaped	01	05	1–5–10–20–36	×	×
Square-shaped	01	05	1–5–10–20–36	×	×
L-shaped	01	05	1–5–10–20–36	×	×
Linear-shaped	01	06	1–5–10–20–40–60	×	×

measuring impulse responses and SPL. Although the source sound, namely a starting clapper for RT and a speaker for SPL attenuation, remained constant in all outside areas, the positions of the reception points, which were microphones, were adjusted along their respective line of sight. The distance between the source and receiver for each measurement zone was obtained by taking into account the size of the outdoor spaces. The source receiver distance was logarithmically scaled within a range of 40 m in four zones. This was done in order to study the distribution of RT and the attenuation of SPL in these outdoor spaces.

A single source position was selected in each courtyard to represent a typical noise source location and to ensure direct comparability between sites. This approach helped isolate the effects of layout, while the enclosed nature of the spaces limited variability from source relocation.

The INR was found to be 22 dB at 125 Hz, 30 dB at 250 Hz, 39 dB at 500 Hz, 51 dB at 1000 Hz, 63 dB at 2000 Hz, 63 dB at 4000 Hz, and 49 dB at 8000 Hz at a distance of 40 m, which is considered to be the greatest distance between the source and the receiver.

To accurately measure RT for T20 and T30, respectively, ISO 3382-2 recommends an INR of at least 35 dB and 45 dB. Based on the INRs in the one band displayed above, the RT calculation method was based on T20 (−5 dB to −25 dB) in one-octave bands from 500 Hz to 8000 Hz for source to receiver distances within 40 m.

In this study, the 125 Hz and 250 Hz octave bands were excluded from the analysis because their impulse-to-noise ratio (INR) values did not meet the minimum threshold required for reliable measurement, as recommended in ISO 3382. This approach is consistent with previous outdoor acoustic studies, including those by Yang *et al.* (2013; 2017), where low-frequency bands were omitted when the INR was insufficient. In outdoor environments, low frequencies are particularly vulnerable to interference from background noise and environmental factors, which can lead to measurement inaccuracies. Therefore, excluding these bands ensured that the reported results were based on robust and reliable data.

3. Results and discussion

3.1. Impulse response

Figure 4 and Fig. 5 provide the pressure squared impulse responses and decay curves recorded at receiver distances of 20 m, in order to verify the impact of building layouts on multiple reflections. In order to analyze

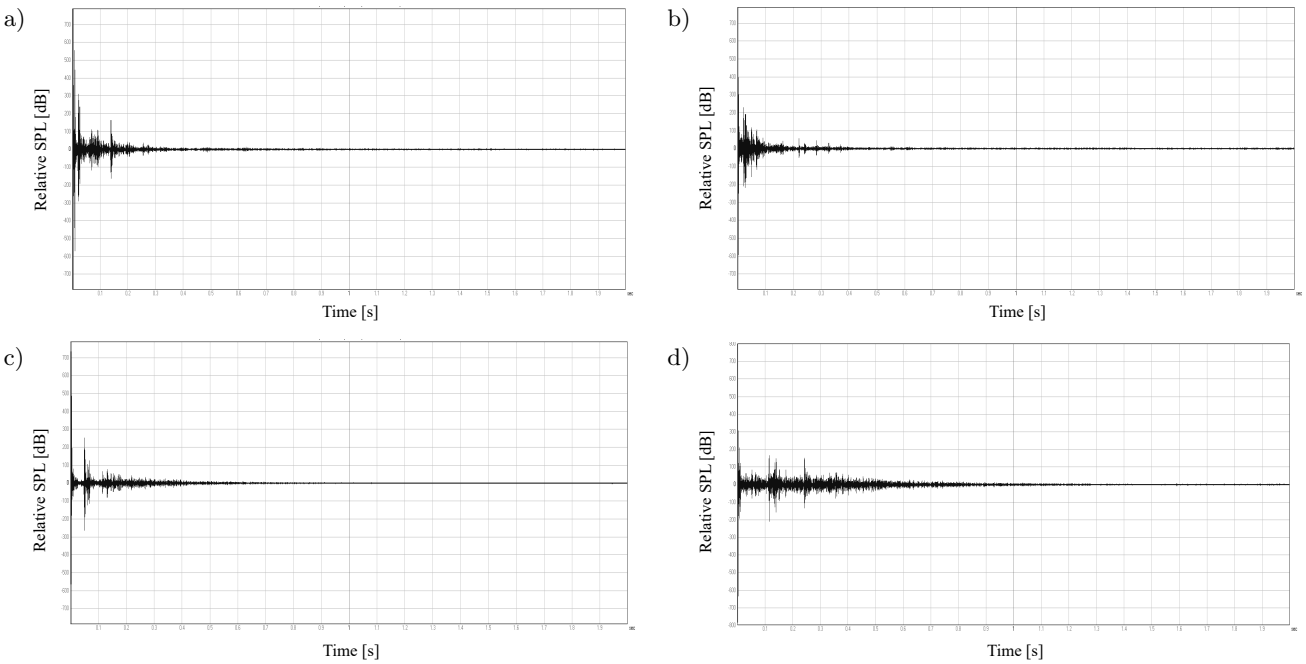

Fig. 4. Impulse responses at 1000 Hz for each of the four outdoor sites measured at a source-to-receiver distance of about 20 m: a) L-shaped, b) linear-shaped, c) U-shaped, d) square-shaped.

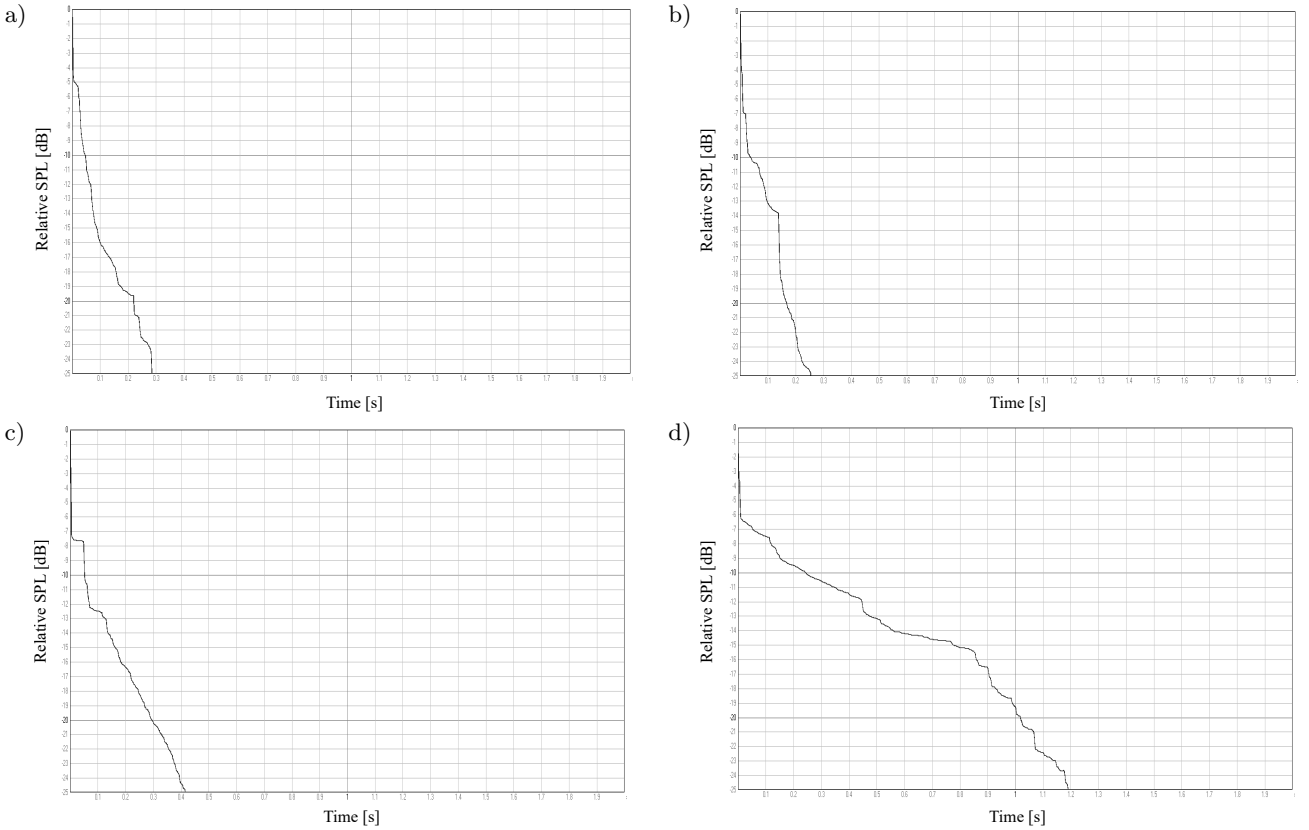

FIG. 5. Decay curves at 1000 Hz for each of the four outdoor sites measured at a source-to-receiver distance of about 20 m: a) L-shaped, b) linear-shaped, c) U-shaped, d) square-shaped.

the variations in sound energy reflection patterns across the four outdoor locations, it is beneficial to compare the impulse responses and decay curves obtained from the same sound source.

The outcome shown in Fig. 4 displays impulse responses that include sound reflections coming later to the direct sound from building façades, ground, and other obstructions. Therefore, it can be concluded that the sound energy that bounces back produces an increase in SPL and RT, which are directly linked to the perception of noise discomfort and spatial impressions. The reflection patterns of impulse responses vary throughout the four outside areas, despite the tests being conducted at equal distances between the source and receiver. The reflection pattern is impacted by several design aspects, including building layout, building form, gaps between buildings, and arrangement of building façades. Based on the various forms of outdoor spaces, the sound energy reflected in U- and square-shaped areas is comparatively strong compared to that in linear- and L-shaped areas. This observation is further supported by the decay curve shown in Fig. 5.

3.2. General features and RT distribution

In this work, RT analysis at low frequencies (125 Hz and 250 Hz) is excluded due to the insufficient INR.

Figure 6 displays the maximum, average, and lowest RT20 values recorded at each measurement area across various frequencies ranging from 500 Hz to 8000 Hz in octave bands. This analysis aims to assess the general features and distribution of RT20 in outdoor environments.

The findings indicate that there is a significant variation in RT20 between the highest and lowest values across different measurement areas, suggesting an uneven distribution of RT20 in the outdoor environment.

The building layout has a significant impact on RT20, as seen by the varying maximum, average, and lowest values of RT20 based on the measurement regions. The RT20 has considerably longer durations at frequencies of 500 Hz, 1000 Hz, and 2000 Hz when compared to other frequencies. The maximum RT20 occurs at a frequency of 1000 Hz for a square-shaped (□) that is longer than 2.38 s.

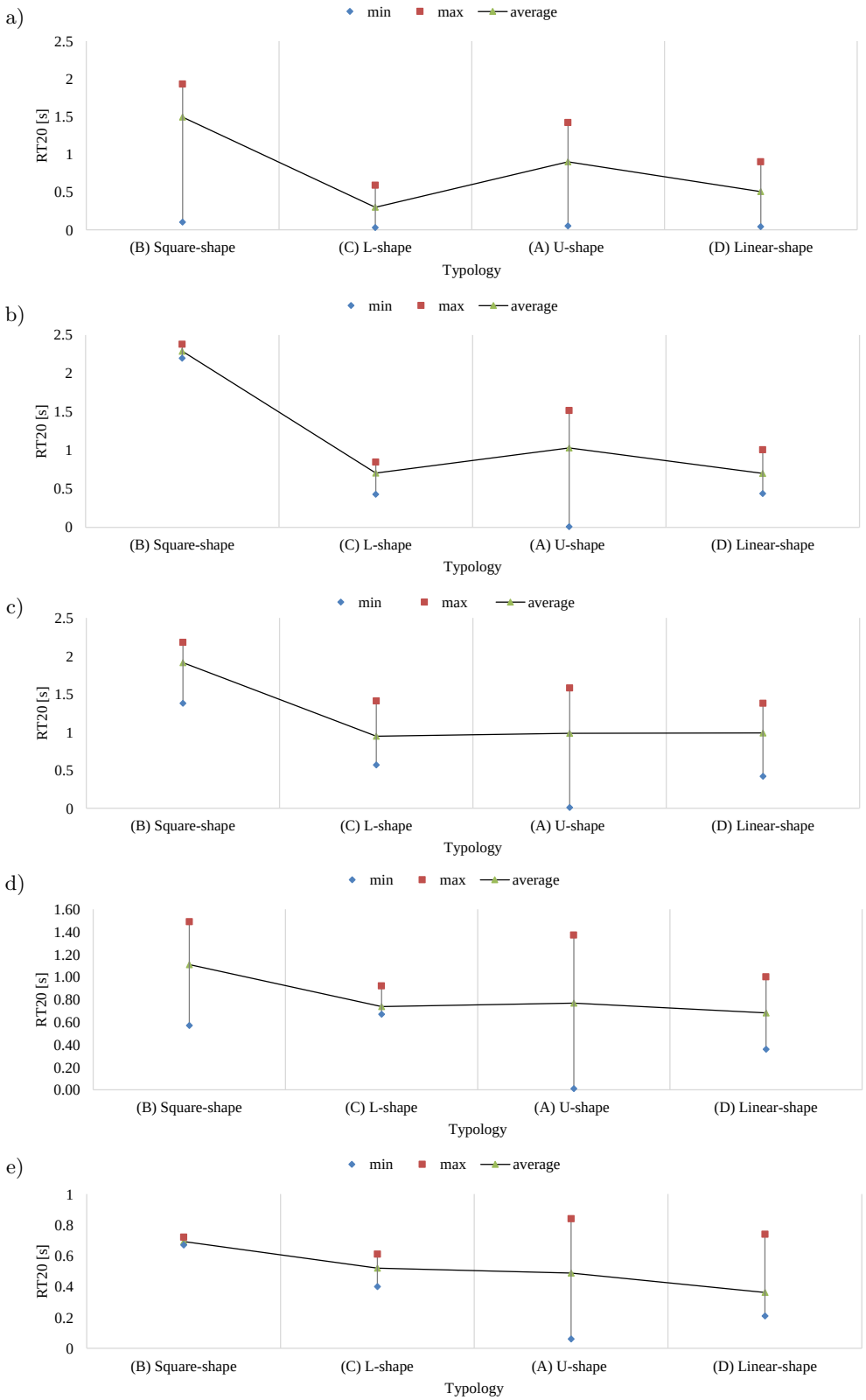

FIG. 6. RT20 values, including the maximum, average, and minimum, with their corresponding frequencies, measured at the four outdoor places at: a) 500 Hz, b) 1000 Hz, c) 2000 Hz, d) 4000 Hz, e) 8000 Hz.

The distance between source and receiver determines RT20 in urban settings. Therefore, the measurement of RT20 at various source receiver distances in the four measurement zones using different sources and receivers were

shown in Fig. 7. Despite the consistent source-receiver distance, the findings demonstrate that there is a significant disparity in RT20 across various measurement zones. This suggests that various architectural designs may have an impact on RT20.

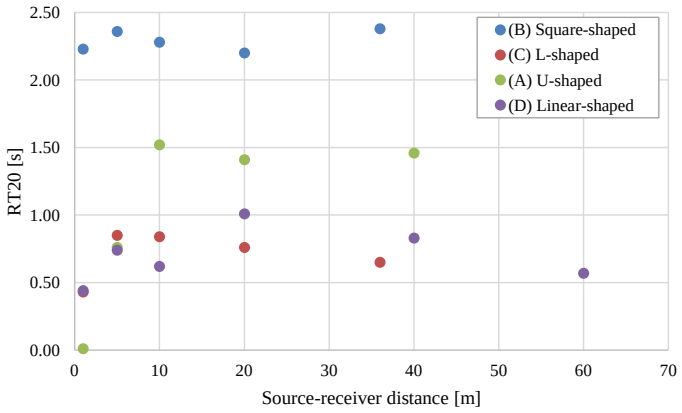

FIG. 7. General RT20 at a frequency of 1000 Hz, measured at four distinct areas using varying source-receiver distances.

In conclusion, the findings referring to the distribution and general features of RT20 suggest that RT20 may be influenced by building layouts and architectural design.

3.3. Acoustic parameters

3.3.1. Reverberation time

RT is one of the main quantitative measures of acoustic parameters that describe the sound behavior in a space. To assess the RT in outdoor spaces, source receiver distance is the determinant factor. Figure 8 displays the observed RT based on the distance between the sound source and receiver for four distinct building layouts. The graph includes regression curves and correlation coefficients (R^2) specifically for the frequency of 500 Hz. The determination of the calculating technique for the regression curve is based on selecting the correlation coefficient with the greatest value. For the polynomial regression curve, the equation of the 2nd order is used.

At 500 Hz, the regression curves across all typologies show a consistent pattern, increasing logarithmically or polynomially with distance. This trend likely results from the decrease in direct sound energy at shorter distances, while the amplitude of reflected sound grows at longer distances. Notably, in the L-shaped configuration, starting from a distance of 20 m, the RT begins to decline with increasing distance, which can be attributed to the reduced effects of reflections. The correlation coefficients (R) observed between these phenomena range from 0.52 to 0.94, signifying a strong correlation.

It is worth noting that the RT values at the same source-receiver distance varied significantly among the (□), (U), (–), and (L) types. For instance, in the (□) type, the RT values were relatively strong compared to those in the (U), (–), and (L) types. This difference can be attributed to the number of façades surrounding the outdoor space. In the (□) type, where the outdoor space is enclosed by a larger number of façades, the sound reflections and reverberations are more pronounced, causing longer RT values. On the other hand, in the (U) and (L) types, which have fewer surrounding façades the sound reflections and reverberations are less prominent, leading to shorter RT values. These variations in RT values demonstrate the significant influence of the architectural design and surrounding structures on the acoustic characteristics of the outdoor space.

The observed variation in RT20 between the different courtyard shapes is consistent with (Yang *et al.*, 2013; 2017), who also found that reverberation characteristics change significantly with architectural configuration. In both studies, more enclosed layouts exhibited higher RT values, while more open configurations showed lower sound persistence.

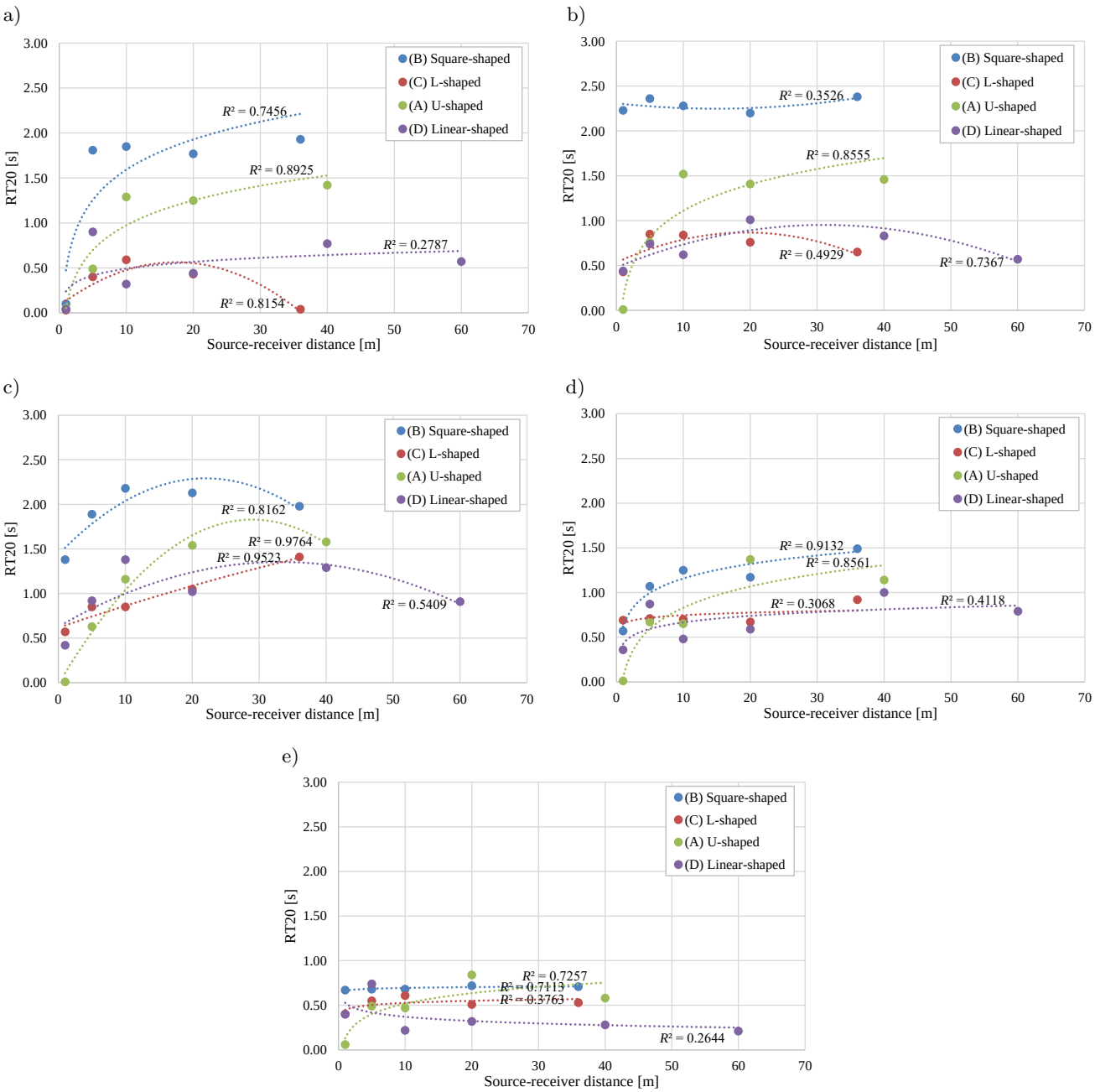

FIG. 8. RT measured based on source–receiver distance for the four different types of building layouts, with regression curves and correlation coefficients R^2 at: a) 500 Hz, b) 1000 Hz, c) 2000 Hz, d) 4000 Hz, e) 8000 Hz.

3.3.2. EARLY DECAY TIME

In Fig. 9, the EDT at each measurement context is shown according to the source to receiver distances. EDT is a parameter extrapolated from the decay curve portion that spans between 0 dB and 10 dB below the initial level. Therefore, the sound energy generation from early reflections affects significantly this parameter. The result in Fig. 9 shows that EDT tends to increase with increasing source to receiver distances (polynomialy), which is similar to RT in all outdoor spaces.

Figure 9 displays the EDT for each measurement context, based on the distances between the sound source and the receiver. EDT is a metric extrapolated from the section of the decay curve that extends from 0 dB to 10 dB below the original level. Consequently, the creation of sound energy from early reflections has a considerable impact on this parameter. The data shown in Fig. 9 demonstrates that the EDT tends to rise in a quadratic

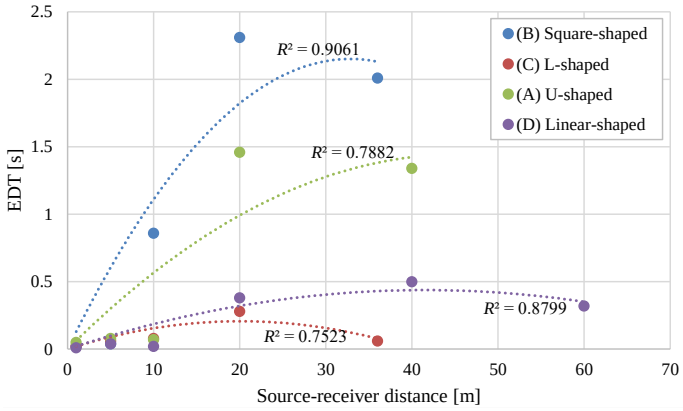

FIG. 9. Measured EDT at 500 Hz with different source to receiver distances for the four different types of building layouts.

manner as the distance between the source and receiver increases. This trend is consistent with the behavior of the RT in all outdoor environments.

Similarly to RT, the correlation coefficient falls within the range of 0.87 to 0.96, signifying a strong and clear correlation between the variables. It can also be seen that at the same source to receiver distance, EDT is similar to RT, it has different values due to the different number of façades surrounding the outdoor space. Understanding such differences is crucial for designing outdoor areas with desired acoustic qualities, whether it is to enhance sound projection and reverberation in performance venues or to ensure speech clarity in public gathering spaces.

The variation in EDT observed across courtyard shapes aligns with the findings of (YANG *et al.*, 2017), where more enclosed configurations yielded longer EDTs due to stronger and more sustained reflections, while open layouts produced shorter EDT values.

3.3.3. DEFINITION (D50)

The clarity of the speech is assessed using D50, a criterion that quantifies the ratio of sound energy arriving within the first 50 ms to the overall sound energy, measured as a percentage.

Figure 10 displays the D50 values at various source receiver distances for the four outdoor areas. In most type spaces, like RT, D50 decrease (polynomial) with the increase of distances. That means that with increasing distances the clarity of sound decreases. Like that of RT, the correlation coefficient of regression curves among these contexts falls within 0.50 and 0.97, indicating strength relationship between variables. At the same distance, despite the D50 is categorized within the range of good to excellent levels, the values differ in each outdoor space, e.g., at 20 m source receiver distance, the D50 is 0.60, 0.61, 0.89, and 0.79 in (□), (U), (L), and

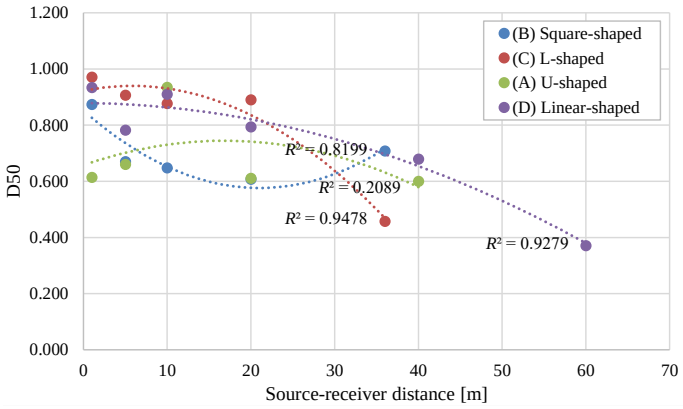

FIG. 10. D50 with different source to receiver distances for the four different types of building layouts.

(–) shape, respectively. This can be attributed to the varying number of façades that encompass the outdoor space. Hence, when designing outdoor spaces, it becomes crucial to consider the distinctive attributes of the building layouts that encircle these open areas.

The differences in D50 values between the various configurations are consistent with the findings of (Yang *et al.*, 2017), which reported that open courtyard forms tend to enhance speech clarity (higher D50) by reducing late reflections, whereas enclosed forms can lower clarity due to increased reverberant energy.

3.3.4. Rapid speech transmission index

The assessment of speech intelligibility in outdoor environments is performed using the RaSTI measure, which takes into account the distance between the sound source and the receiver. The evaluation is determined by five unique levels, with each level corresponding to a particular range; 0–0.3 is classified as extremely bad, 0–0.45 as poor, 0.45–0.6 as fair, 0.6–0.75 as good, and 0.75–1.0 as exceptional (International Electrotechnical Commission, 2020).

According to the data shown in Fig. 11, the RaSTI generally decreases as the distance increases in most typology settings. This trend is comparable to the findings of D50. The reason for this is because while the distance between the source and receiver is small, the direct sound has a greater influence on the initial sound energy of the impulse response, leading to a shorter RT. However, as the distance rises, the amplitude of the direct sound decreases, causing the RT to increase.

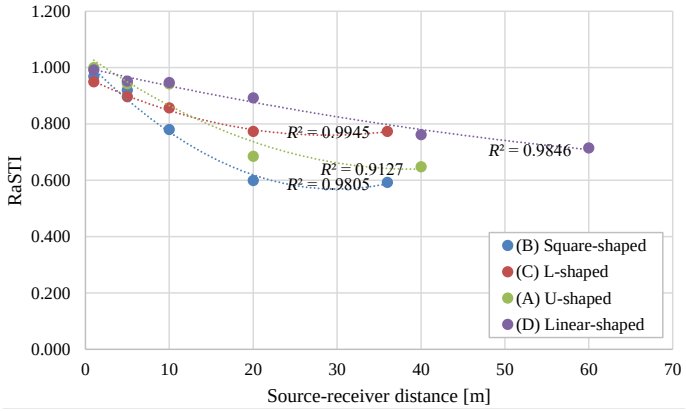

Fig. 11. RaSTI with different source to receiver distances for the four different types of building layouts.

At the same distance, despite the RaSTI is characterized within the range of good to excellent, the values vary in each outdoor space. For example, at the 20 m source receiver distance, the RaSTI value is 0.6, 0.68, 0.77, 0.89 in (□), (U), (L), and (–) shape, respectively. This is because of the different number of façades surrounding the outdoor space. Hence, the design of outdoor spaces must take into account the attributes of the building layouts that surround the outdoor area.

The measured changes in RaSTI with different courtyard shapes correspond with the findings of (Yang *et al.*, 2017), indicating that open configurations generally improve speech intelligibility, while more enclosed geometries may limit it due to prolonged reverberation and multiple reflection paths.

3.3.5. Sound pressure level

Figure 12 presents the SPL attenuation results in comparison with the reference SPL, which was obtained at a distance of 1 meter between the source and receiver, in five outdoor areas. To interpret the results, the measurements were compared with the semi-free field attenuation, where SPLs are expected to decrease by approximately 6 dB each time the distance from the source doubles in an unobstructed environment. This provided a reference baseline to evaluate how the presence of surrounding building façades and courtyard configurations modified sound propagation in the studied outdoor spaces. The findings indicate that SPL diminishes as the distance between the source and receiver increases in all outdoor areas, owing to the properties of the non-diffuse field.

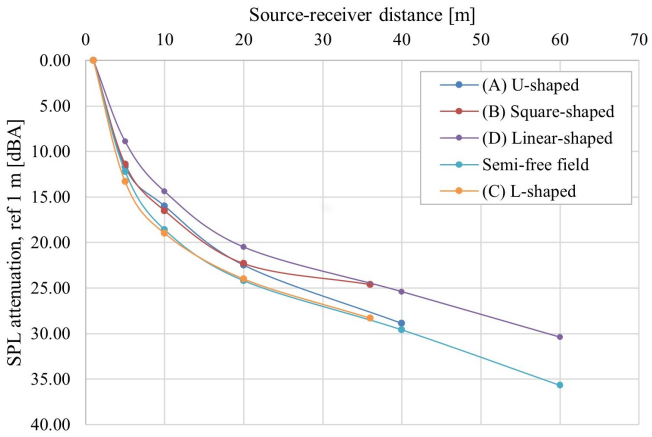

Fig. 12. SPL attenuation according to source to receiver distance.

Within a distance of 1 m to 5 m between the source and receiver, there is no notable variation in SPL reduction across the five outdoor areas. This is because the direct sound plays a prominent role.

However, in the far field, at the same position where the sound source and receiver are located, it can also be seen that the SPL decreases differently depending on the outdoor arrangement and the features of the surrounding geometry. Although the linear-shape is surrounded by one side of building façade, it shows the lowest SPL attenuation. This is because the high sound reflections off surfaces such as bitumen and pavement ground. The SPL attenuations in square- and U-shaped outdoor spaces are similar within the source-receiver distance of 10–20, with square-shaped space exhibiting lower attenuation beyond that distance. This difference occurs because U-shaped spaces allow for less reflection energy compared to square-shaped ones. The highest SPL attenuation is revealed in the L-shaped outdoor spaces showing a similarity with SPL attenuation in the semi free field. This is due to the lack of reflections toward the outdoor space. The overall outcome suggests that the architecture of the building layout has a substantial impact on the degree of noise irritation that students feel.

The SPL attenuation patterns match findings from (Yang *et al.*, 2017), showing that open layouts allow sound to disperse more rapidly, leading to higher attenuation rates, while enclosed layouts slow down attenuation due to boundary reflections and energy confinement.

4. Conclusion

The present research conducted a series of field measurements to assess SPL attenuation and room acoustical parameters including RT, EDT, RaSTI, and D50 across four outdoor spaces within the University of Batna 1. These spaces were chosen to represent diverse building layouts and blocks.

Overall, variations in the maximum, average, and minimum values of RT20 were noted across different measurement areas, underscoring the influence of building layout on RT20 distribution. The square shape retains sound the longest, the U-shape shows the most irregular reverberation, and the linear shape is the most stable. At high frequencies, all layouts have faster sound decay and less variability due to greater absorption and scattering.

Similarly, as the distance between the sound source and receiver increased, the findings indicated that architectural design substantially influences the dispersion of acoustic energy. Among the many layouts examined (square, U, L, and linear) – the square- and U-shaped courtyards had the highest RT20 values, especially at mid-range frequencies (1000 Hz to 2000 Hz), attributable to their enclosed geometry that captures sound energy. The linear- and L-shaped courtyards promoted expedited sound fading, demonstrating reduced RT20 values and enhanced SPL attenuation stability, rendering them more appropriate for settings necessitating improved speech intelligibility.

The impulse response research verified that U-shaped and square layouts produced more intense reflections, resulting in extended reverberation, whilst the linear arrangement had the lowest RT20 values, indicating effective sound dissipation.

The D50 and RaSTI values were greatest in linear courtyards, indicating enhanced speech intelligibility owing to less reverberation and increased direct sound prominence. In contrast, square- and U-shaped courtyards had reduced D50 and RaSTI values, indicating diminished speech intelligibility resulting from extended reverberation.

The SPL attenuation research indicated that linear courtyards had a more steady decline in SPL with distance, whereas square- and U-shaped areas displayed variable SPL patterns, affected by heightened reflections inside their confined perimeters.

The results of this study highlight that building layout plays a decisive role in shaping the outdoor acoustic environment, with measurable impacts on both sound persistence (RT20, EDT) and speech intelligibility (D50, RaSTI). This knowledge can inform architectural and urban design decisions, particularly when planning courtyards, campus spaces, and other semi-enclosed outdoor areas.

For instance, layouts with more enclosed geometries, such as square- and U-shaped forms, may be preferred in contexts where sound retention is desirable – such as cultural performances or ceremonial events – due to their capacity to preserve acoustic energy. Conversely, linear- and L-shaped configurations, which promote quicker sound dissipation and higher speech clarity, may be better suited for everyday circulation spaces, recreational zones, or public areas where speech intelligibility and noise reduction are priorities.

Furthermore, these findings provide a basis for integrating acoustic considerations early in the spatial planning process, alongside visual, thermal, and functional criteria. By anticipating how form and enclosure affect sound propagation, designers can create outdoor environments that are acoustically tailored to their intended uses.

Finally, while this study focuses on a specific university setting, the principles identified are broadly applicable to urban courtyards, plazas, and pedestrian streets. Future research could extend this work by incorporating variations in building height, façade material properties, vegetation, and seasonal changes to develop comprehensive design guidelines for acoustically optimized outdoor spaces.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Authors' contributions

Sami Hamouta conceptualized the study, performed the analysis, interpreted the data, and wrote the original draft. Atef Ahriz and Noureddine Zemmouri conceptualized the study, contributed to data analysis and interpretation, and wrote the original draft. Ahmed Mansouri contributed to data analysis and wrote the original draft. All authors reviewed and approved the final manuscript.

Acknowledgments

The authors would like to express their gratitude to the following organizations and individuals for their valuable contributions to this research:

– Ahnert Feistel Media Group (AFMG) company for generously providing a free version of EASEA Pro software, which was instrumental in conducting the measurements and analyses for this research. This software significantly contributed to the outcomes of our study, and we appreciate AFMG's support in facilitating our work.

– Laboratory of Child, City and Environment (LEVE) directed by Professor Dib Belkacem, for generously providing equipment for conducting field measurement.
– M.C. Mansouri, A. Mestiri, S. Haddad, and T. Benfifi for providing assistance with data collection.

We also extend our appreciation to the faculty and staff of the University of Batna 1 for their support and cooperation throughout this study.

References

1. Ariza-Villaverde A.B., Jiménez-Hornero F.J., Gutiérrez De Ravé E. (2014), Influence of urban morphology on total noise pollution: Multifractal description, *Science of the Total Environment*, **472**: 1–8, https://doi.org/10.1016/j.scitotenv.2013.10.091.
2. Aylor D., Parlange J.-Y., Chapman C. (1973), Reverberation in a city street, *The Journal of the Acoustical Society of America*, **54**(6): 1754–1757, https://doi.org/10.1121/1.1914476.
3. Benameur O., Zemmouri N., Cutini V., Leccese F., Salvadori G. (2022), Exploration of environmental noise in Saharan oases on the basis of urban configurations: City of Biskra datasets, *Data in Brief*, **43**: 108392, https://doi.org/10.1016/j.dib.2022.108392.
4. Bouzir T.A.K., Zemmouri N. (2017), Effect of urban morphology on road noise distribution, *Energy Procedia*, **119**: 376–385, https://doi.org/10.1016/j.egypro.2017.07.121.
5. Çolakkadioğlu D., Yücel M., Kahveci B., Aydinol Ö. (2018), Determination of noise pollution on university campuses: A case study at Çukurova University campus in Turkey, *Environmental Monitoring and Assessment*, **190**(4): 203, https://doi.org/10.1007/s10661-018-6568-8.
6. Echevarria Sanchez G.M., Van Renterghem T., Thomas P., Botteldooren D. (2016), The effect of street canyon design on traffic noise exposure along roads, *Building and Environment*, **97**: 96–110, https://doi.org/10.1016/j.buildenv.2015.11.033.
7. Eggenschwiler K., Heutschi K., Taghipour A., Pieren R., Gisladottir A., Schäffer B. (2022), Urban design of inner courtyards and road traffic noise: Influence of façade characteristics and building orientation on perceived noise annoyance, *Building and Environment*, **224**: 109526, https://doi.org/10.1016/j.buildenv.2022.109526.
8. Flores R., Gagliardi P., Asensio C., Licitra G. (2017), A case study of the influence of urban morphology on aircraft noise, *Acoustics Australia*, **45**(2): 389–401, https://doi.org/10.1007/s40857-017-0102-y.
9. Goines L., Hagler L. (2007), Noise pollution: A modem plague, *Southern Medical Journal*, **100**(3): 287–294, https://doi.org/10.1097/smj.0b013e3180318be5.
10. Goswami S., Nayak S.K., Pradhan A.C., Dey S.K. (2011), A study on traffic noise of two campuses of University, Balasore, India, *Journal of Environmental Biology*, **32**(1): 105–109.
11. Gulwadi G.B., Mishchenko E.D., Hallowell G., Alves S., Kennedy M. (2019), The restorative potential of a university campus: Objective greenness and student perceptions in Turkey and the United States, *Landscape and Urban Planning*, **187**: 36–46, https://doi.org/10.1016/j.landurbplan.2019.03.003.
12. Han X., Huang X., Liang H., Ma S., Gong J. (2018), Analysis of the relationships between environmental noise and urban morphology, *Environmental Pollution*, **233**: 755–763, https://doi.org/10.1016/j.envpol.2017.10.126.
13. International Electrotechnical Commission (2020), *Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index* (IEC Standard no. IEC 60268-16:2020), https://webstore.iec.ch/publication/26771.
14. Lee P.J., Kang J. (2015), Effect of height-to-width ratio on the sound propagation in urban streets, *Acta Acustica*, **101**(1), https://doi.org/10.3813/AAA.918806.
15. Montalvão Guedes I.C., Bertoli S.R., Zannin P.H.T. (2011), Influence of urban shapes on environmental noise: A case study in Aracaju – Brazil, *Science of the Total Environment*, **412–413**: 66–76, https://doi.org/10.1016/j.scitotenv.2011.10.018.
16. Oliveira M.F., Silva L.T. (2011), The influence of urban form on facades noise levels, [in:] *WSEAS Transactions on Environment and Development*, **7**(5).

17. Picaut J., Le Pollčs T., L'Hermite P., Gary V. (2005), Experimental study of sound propagation in a street, *Applied Acoustics*, **66**(2): 149–173, https://doi.org/10.1016/j.apacoust.2004.07.014.

18. Silva L.T., Oliveira M., Silva J.F. (2014), Urban form indicators as proxy on the noise exposure of buildings, *Applied Acoustics*, **76**: 366–376, https://doi.org/10.1016/j.apacoust.2013.07.027.

19. Steenackers P., Myncke H., Cops A. (1978), Reverberation in town streets, *Acta Acustica United with Acustica*, **40**(2): 115–119.

20. Su W., Kang J., Jin H. (2013), Acoustic environment of University Campuses in China, *Acta Acustica*, **99**(3), https://doi.org/10.3813/AAA.918622.

21. Thomas P., Van Renterghem T., De Boeck E., Dragonetti L., Botteldooren D. (2013), Reverberation-based urban street sound level prediction, *The Journal of the Acoustical Society of America*, **133**(6): 3929–3939, https://doi.org/10.1121/1.4802641.

22. Wang B., Kang J. (2011), Effects of urban morphology on the traffic noise distribution through noise mapping: A comparative study between UK and China, *Applied Acoustics*, **72**(8): 556–568, https://doi.org/10.1016/j.apacoust.2011.01.011.

23. Wang L.K., Pereira N.C., Hung Y.-T. (2005), *Advanced Air and Noise Pollution Control*, Humana Press, https://doi.org/10.1007/978-1-59259-779-6.

24. Wiener F.M., Malme C.I., Gogos C.M. (1965), Sound propagation in urban areas, *The Journal of the Acoustical Society of America*, **37**(4): 738–747, https://doi.org/10.1121/1.1909409.

25. World Health Organization (2018), *Environmental noise guidelines for the European region*, https://www.who.int/europe/publications/i/item/9789289053563 (access: 15.08.2023).

26. Xie H., Kang J., Tompsett R. (2011), The impacts of environmental noise on the academic achievements of secondary school students in Greater London, *Applied Acoustics*, **72**(8): 551–555, https://doi.org/10.1016/j.apacoust.2010.10.013.

27. Yang H.-S., Kang J., Kim M.-J. (2017), An experimental study on the acoustic characteristics of outdoor spaces surrounded by multi-residential buildings, *Applied Acoustics*, **127**: 147–159, https://doi.org/10.1016/j.apacoust.2017.05.037.

28. Yang H.-S., Kim M.-J., Kang J. (2013), Acoustic characteristics of outdoor spaces in an apartment complex, *Noise Control Engineering Journal*, **61**(1): 1–10, https://doi.org/10.3397/1.3702001.

29. Yeow K.W. (1977), Decay of sound levels with distance from a steady source observed in a built-up area, *Journal of Sound and Vibration*, **52**(1): 151–154, https://doi.org/10.1016/0022-460X(77)90399-6.

30. Zannin P.H.T., Engel M.S., Fiedler P.E.K., Bunn F. (2013), Characterization of environmental noise based on noise measurements, noise mapping and interviews: A case study at a university campus in Brazil, *Cities*, **31**: 317–327, https://doi.org/10.1016/j.cities.2012.09.008.

31. Zannin P.H.T., Ferraz F. (2016), Assessment of indoor and outdoor noise pollution at a university hospital based on acoustic measurements and noise mapping, *Open Journal of Acoustics*, **6**(4): 71–85, https://doi.org/10.4236/oja.2016.64006.

32. Zannin P.H.T., Zwirtes D.P.Z. (2009), Evaluation of the acoustic performance of classrooms in public schools, *Applied Acoustics*, **70**(4): 626–635, https://doi.org/10.1016/j.apacoust.2008.06.007.

33. Zuccherini Martello N., Fausti P., Santoni A., Secchi S. (2015), The use of sound absorbing shading systems for the attenuation of noise on building façades. An experimental investigation, *Buildings*, **5**(4): 1346–1360, https://doi.org/10.3390/buildings5041346.

Archives of Acoustics Vol. 51, No. 2, pp. 317–331 (2026), https://doi.org/10.24423/archacoust.2026.4328

Research Paper

The Influence of the Surface of Ventilation Duct on Sound Attenuation in the Airflow

Joanna Maria Kopania[(1)*], Kamil Wójciak[(1)], Patryk Gaj[(1)], Grzegorz Boguławski[(2)]

[(1)] *Institute of Power Engineering – National Research Institute*
Warsaw, Poland

[(2)] *Lodz University of Technology*
Lodz, Poland

*Corresponding Author: joanna.kopania@p.lodz.pl

Received September 8, 2025; revised January 8, 2026; accepted March 11, 2026; available online March 31, 2026; version of record June 3, 2026; published issue June 24, 2026.

These studies focus on acoustical parameters of steel flat-oval ducts as a function of their roughness. The four types of steel ducts were measured: raw steel, galvanised steel, painted steel, and aluminium as the reference one. The roughness of the duct was measured, and roughness parameters were specified. The sound power level was obtained on the specially constructed stand test with an outlet to the reverberation room. Insertion losses to evaluate the acoustic attenuation performance of the studied steel ducts were obtained. In the present study, an aluminium duct, which is very smooth with minimal airflow friction, was treated as a low-noise object ('silencer'). These studies have shown that for each of the tested steel ducts, the self-noise is higher than for the aluminium duct. The largest differences in this self-noise were observed at a velocity of 12 m/s for the galvanised duct and the raw steel duct compared to the aluminium duct. Insertion losses in straight ducts are consistent with literature and are very low for flat-oval steel ducts. Aluminium duct performs better acoustically than the other ducts studied at lower velocities; however, as airflow velocity increases, the differences in acoustic performance between the materials become less pronounced. This suggests that aerodynamic effects dominate over material surface treatments at higher velocities.

Keywords: steel duct, roughness, insertion loss, HVAC.

Notations

D_i – insertion loss,
L_W – sound power level,
L_{WI} – sound power level when the test object,
L_{WII} – sound power level references object,
P – duct perimeter,
R_a – average roughness,
R_q – root mean square (RMS) roughness,
R_z – average maximum height of profile,
S – area ratio,
v – flow velocity,
ΔL_w – differences between single-number values of sound power level.

1. Introduction

Most standards and guidelines for the analysis of sound in ducts are based on the sound power-based description, which has been widely utilised in HVAC systems, for example, ASHRAE (Reynolds, Bledsoe, 1991; American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2007), VDI (VDI, 2001), or ISO5136 (International Organization for Standardization, 2003), however, it cannot be applied for frequencies above the

cut-off frequency. The sum of the sound power level gains and losses at the component connecting interfaces from the fan to the network's terminal sections can also be used to calculate the sound power level inside a duct network. However, this method ignores any influences caused by wave reflections. This significantly lowers the reliability while simplifying the forecast process. Additionally, the low-frequency range is covered by the plane wave-based description of networks, ducts, and mufflers (Munjal, 1987; Boden, Abom, 1995; Boden, Glav, 2007). However, in HVAC duct networks with their large lateral dimensions the frequency range of plane wave propagation is rather limited. If the duct system is fairly extensive, then existing prediction methods allow the design of HVAC systems that are usually free from noise problems.

The source of the duct break-out in ventilation systems can be calculated thanks to the several prediction algorithms utilised in the ASHRAE scheme. There are several regions where noise is produced, such as diffuser noise, system attenuation from plenum chambers, unlined rectangular ducts and bends, end reflection losses of ducts, breakout from ducts under airflow conditions, and insertion losses of ducts (Reylnods, Bledsoe, 1991). The Allen formula (Allen, 1960) is a well-known technique for forecasting noise levels that emerges from a section of a ductwork. The difference between the areas through which sound enters the duct (i.e., its cross sectional area) and exits the duct (i.e., the duct surface area) is taken into account. It is based on the theory of sound transmission through panels. However, it is widely acknowledged that this approach occasionally produces incredibly erroneous results, especially for low frequencies and lengthy ducts. The approach fails to consider the fact that the sound power within the duct will decrease along its length as a result of breakout from the duct. Negative attenuations may be predicted using the formula, which means that more sound energy is emitted from the duct than enters it. This is due to the fact that using the sound reduction index (or transmission loss) data for duct walls was gathered from sound transmission loss measurements on plane panels rather than in duct tests, presents another challenge. From this time, various analytical and numerical methods have been employed in the modelling of duct wall break-out and break-in (Cummings, 1980; 1983; Cummings *et al.*, 1984; Venkatesham *et al.*, 2011; Herrin, Seybert, 2006). However, because of the various types of noise sources and the various ways that noise spreads, evaluating the noise produced by HVAC systems can be challenging. It takes a lot of programming work to implement any accurate predictive model for duct wall break-out and break-in because the self-noise produced by aerodynamic sources through duct systems depends on the actual system configuration and the effective airflow velocity, although we are also dealing with sound attenuation of various devices and ducts in the whole system. In turns, the attenuation in HVAC systems is a complicated issue.

Sound energy splits at branches, gets rejected at bends and duct terminations, and loses energy due to duct wall vibration, all contributing to the natural attenuation provided by the duct system. The attenuation caused by the duct wall vibration may become significant when the system has extensive ductwork, especially for rectangular ducts. The produced motion of the duct walls reduces the energy of a sound wave as it travels down an unlined duct. The wall mass is mostly responsible for the surface impedance, and the computation of the duct loss is similar to that of the transmission loss. It is significantly harder to excite circular sheet metal ducts than rectangular ones at low frequencies, especially when they are in their first state of vibration, known as the breathing mode. Consequently, unlined rectangular ducts suppress sound significantly more than circular ducts (Cummings, 2001).

Empirical equations for the attenuation was development by Reynolds and Bledsoe (1991) in terms of a duct perimeter (P) to area ratio (S). A large P/S ratio means that the duct is wide in one dimension and narrow in the other. The attenuation of rectangular ducts in the 63 Hz to 250 Hz octave frequency bands can be approximated by using an equation (Reynolds, Bledsoe, 1991):

– for $\frac{P}{S} \geq 3$:

$$D = 17.0 \times \left(\frac{P}{S}\right)^{-0.25} \times f^{-0.85} \times L, \tag{1}$$

– for $\frac{P}{S} < 3$:

$$D = 1.64 \times \left(\frac{P}{S}\right)^{0.73} \times f^{-0.58} \times L. \tag{2}$$

The attenuation of rectangular ducts above 250 Hz octave frequency bands can be approximated by using:

$$D = 0.02 \times \left(\frac{P}{S}\right)^{0.8} \times L. \tag{3}$$

In these formulas, pay attention to the units of measurement: the value of P must be given in feet, the value of A (the area of the duct) in square meters, and the value of L (the length of the duct) in feet. Ductwork in industrial installations can be several dozen or even several hundred meters long. Acoustic losses in the air on the duct walls can dampen the sound transmitted to the duct and its surroundings, which travels great distances. For steel ducts of 200 mm in diameter in ventilation systems, there is a loss of 0.1 dB/m for low frequencies and 0.3 dB/m for high frequencies, as shown by Bessac *et al.* (2018). For lengthy ducts, these tiny values can result in notable attenuations. Experience has shown that design contractors do not pay attention to this loss and frequently overlook this occurrence in their calculations. The AcouReVe project carried out, between 2015–2018, aimed to improve the reliability and quality of acoustic calculations in the ventilation ductwork and provide more reliable input data and insights for acoustical consultants to predict sound levels in rooms more accurately when considering ventilation systems and ductwork. For straight round ducts, it was confirmed that for low frequencies, the measured losses were consistent with literature, showing very low attenuation (~0.1 dB/m). However, for high frequencies (above 1600 Hz), galvanised spiral steel ducts exhibited higher losses than expected, ranging from 0.5 dB/m to 1.5 dB/m, and even up to 3 dB/m in some specific test cases, contradicting the common assumption of minimal losses in circular ducts (Bessac *et al.*, 2018).

The ducts that run to supply registers are frequently built of flexible aluminium or Mylar or sheet metal. Duct surfaces should ideally be maintained dry and clean; however, even brand-new ducts can get dirty from debris from the building's construction and storage before installation. Furthermore, residual oils from the initial machining and construction of new steel ducts have been found to be sources of volatile organic compounds (VOCs) in indoor air (Pasanen *et al.*, 1995). This is connected to the fact that the commercial pipes have different levels of roughness, which affects the behaviour of the fluid flowing through the pipe, particularly the pressure loss caused by friction. There is a strict connection between roughened pipes and the generation of noise. Roughness can influence fluid flow and heat transfer by increasing pressure drop, altering flow regime laminar-turbulent transition, and by inducing secondary flow motions which lead to an enhancement in flow mixing and heat transfer. Also noise radiated from the exhaust of two roughened pipes to the anechoic chamber was studied by Hersh (1983). He claimed that over the whole tested frequency range (down to roughly 3 kHz), totally rough surfaces generated noticeably more sound. The formation of sound can be explained by the scattering on the surface imperfections, which changed turbulent near-field pressure fluctuations (Howe, 1988; 1998). According to Devenport *et al.* (2011), the flow at the rough wall creates quantifiable roughness noise. It was discovered that even hydrodynamically smooth roughness generated noise, which means that the scattering mechanism was a noise source. As the flow speed increases, the wideband roughness noise moves to a higher frequency and rises in level. The roughness noise is also significantly impacted by the roughness size. These changes with the flow speed and roughness size were found to be incompatible with any straightforward scaling. The presence of roughness of pipe wall changes their values in comparison to the smooth design and, as a result, alters the acoustic behaviour. The Monte Carlo method and the finite element solver were used by Yu *et al.* (2024) to analyse the attenuation of acoustic waves in rough cylindrical pipes and to estimate the roughness of the pipes from the measured acoustic wave. However, there is evidence in numerous studies that the research on noise in HVAC systems should be expanded to include a greater variety of noise sources, as well as the interaction between the different components of the duct system and its impact on noise generation and attenuation (Fry, 1988), especially where the roughness is taken into account. This is because the prediction of system attenuation is usually rather conservative, so that predicted sound pressure levels (SPLs) tend to be overestimated, thus in effect incorporating a factor of safety in the design (Henson, 1986).

The purpose of this paper is to clarify the effect of the surface of the steel duct on the aerodynamic sound and also to determine the attenuation of these ducts, which may be essential for engineering design. The experiments were performed on the test stand, meeting the requirements of ISO 7235, which allows to determine the sound

power level using the precise method. In these studies, the three steel ducts were used compared to the aluminium one as the base channel. The roughness of the channels was also measured.

2. Test objects

The most important parameters of ducts used in HVAC systems are diameter, shape, and material. In ventilation systems, ducts are most commonly constructed from sheet metal, galvanised steel sheets, stainless steel sheets, aluminium, and PVC pipes. Ducts can also be made of aluminium foil or plastic. The choice of material is usually determined by price and design requirements. Steel ducts, often galvanised, are used in industrial ventilation solutions. These systems utilise rectangular or round ducts and, less frequently, flat-oval ducts.

Flat oval ducts are becoming increasingly popular in HVAC systems, as this configuration can enhance airflow dynamics and reduce material consumption. Botejara-Antúnez *et al.* (2023) emphasized the importance of an efficient duct design in minimizing the embodied carbon of HVAC systems, suggesting that careful consideration of the duct shape can lead to both environmental and economic benefits. Studies on flat-oval ducts underline their effectiveness in optimizing airflow patterns, as illustrated in numerical investigations that demonstrate the impact of duct shape on thermal-flow characteristics (Djeffal *et al.*, 2021; Tahrour *et al.*, 2022). The aerodynamic performance of these ducts can lead to better energy efficiency within HVAC applications, as exemplified by Tahrour *et al.* (2022), which evaluated various tube shapes, including flat-oval, for their performance in heat exchangers.

This study utilised flat-oval ducts with large side dimensions (500 mm by 800 mm), specifically designed for industrial HVAC installations – Fig. 1. The ducts were made of raw sheet metal, galvanised sheet metal, and painted sheet metal with two coats of acrylic paint. Additionally, a drawn aluminium duct was used as the base duct. These ducts are characterised by varying surface roughness. The roughness of steel duct (pipe) refers to an index of surface quality, which is usually used to describe the smoothness of the surface of steel duct after processing. The surface roughness of ducts significantly affects their performance. The excessive surface roughness of ducts will affect the resistance, vortex, and friction of the fluid, reducing the transportation efficiency and the stability of fluid flow. Roughness is usually expressed in R_a value (average roughness), R_q value (root mean square (RMS) roughness) or R_z (average maximum height of profile). In this case, the roughness of the inside pipe wall of all pipes was measured with the surface roughness tester 502.300. The instrument gave as output roughness parameters (R_a, R_q, R_z). The roughness parameters were measured at the bottom of the duct in the axial direction. Each measurement was performed at least three times across three different points of the duct, and the results are presented as the average of these measurements. The uncertainty budget for the R_a, R_q, R_z measurements provides a comprehensive breakdown of the potential errors associated with the surface roughness test. The research took into account uncertainties in type A, derived from statistical repeatability across nine measurements, and type B, which includes the instrument's inherent technical limitations. In this case, the primary source of uncertainty was the instrument error, contributing 5 % uncertainty based on manufacturer specification and a 3 % repeatability factor, which accounted for the surface's spatial variability. By combining these factors geometrically, the combined standard uncertainty is obtained. The coverage factor of $k = 2$ gave the expanded uncertainty, ensuring a 95 % confidence level for the reported value. The dimensions of ducts and parameters R_a, R_q, and R_z with uncertainty are given in Table 1.

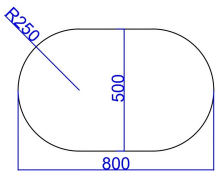

Fig. 1. 2D model of the studied ducts.

The raw steel sheet has a very rough surface with large irregularities. Probably high airflow resistance (turbulent flow) could be observed. Additionally, the surface is poor for sealing or paint adhesion and could trap dust

TABLE 1. Characteristic parameters of the studied ducts.

No.	Duct	Dimension [mm × mm]	Thickness [mm]	Surface condition	R_a [µm]	R_q [µm]	R_z [µm]
1	Raw steel sheet	500 × 800	3	Hot-rolled	12.50 ±0.76	13.75 ±0.84	50.00 ±3.06
2	Galvanised steel sheet	500 × 800	3	Mill finish (as-rolled)	6.30 ±0.39	6.93 ±0.42	25.20 ±1.50
3	Painted steel sheet	500 × 800	3	High-gloss, baked acrylic	1.60 ±0.10	1.76 ±0.11	6.40 ±0.39
4	Drawn aluminium	500 × 800	3	Smooth, seamless	0.80 ±0.05	0.88 ±0.06	2.00 ±0.12

and moisture, which leads to the corrosion risk. The galvanised steel sheet has a moderately rough surface texture; it is better than the raw steel, but still noticeable. The higher airflow resistance than with smooth surfaces could be expected. Suitable for general ducting, but this is not an ideal solution when low drag or ease of cleaning is critical. The painted steel has a smooth surface and low friction in airflow. The surface is mostly smooth, with no sharp defects. But R_z shows some moderate peaks/valleys, but likely not enough to cause airflow turbulence or sealing issues. This duct would be good for cleanrooms or visual/architectural applications. The aluminium duct is very smooth, almost polished, which causes minimal airflow friction. It is excellent for clean environments, lab ducts, or aerospace-grade HVAC. It can be assumed that this channel offers high performance for sealing and corrosion resistance.

3. EXPERIMENT

The experiments were performed on the specific test stand constructed in accordance with two standards PN-EN 3741:2011 'Determination of sound power levels of noise sources using sound pressure – Precision methods for reverberation rooms' and PN-EN 7235:2009 'Acoustics. Laboratory measurement procedures for ducted silencers and air-terminal units. Insertion loss, flow noise and total pressure loss' (see Fig. 2).

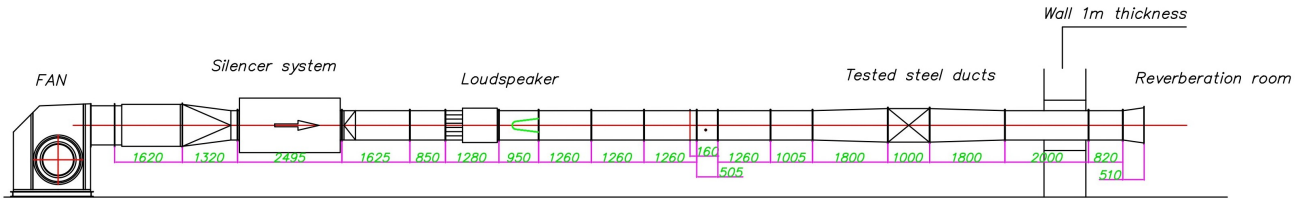

FIG. 2. Test stand with outlet to the reverberation room.

The measurement stand consists of a two-axis fan, a straight duct with a transition to a system of three absorption-resonance silencers, two straight ducts, one of which has an airflow straightener, a straight duct, a loudspeaker chamber, a system of straight ducts (6.5 m), a transition to the test duct, the test duct, transition to the measuring station, and a system of straight ducts with an outlet to the reverberation chamber (3.2 m). The dimensions of the station are shown in Fig. 1. Individual sections of the measuring station have flanges with a glued seal for screw mounting. Thanks to stand test construction, it is possible to determine the sound noise level of devices operating in the flow together with their attenuation effect by the use of precise methods. The reverberation chamber has a volume of 237.0 m^3 and an area of 231.5 m^2, with non-parallel, sound-reflecting walls.

The reverberation chamber has an additional door in the upper part of one of the side walls that opens into a compensation space connected to the laboratory's ventilation system, ensuring free air intake from the chamber and pressure equalisation. Since the outlet of the measuring station is located right next to the wall of the reverberation chamber, the airflow in the chamber space undergoes free expansion due to its dimensions, which causes laminar outflow of excess air from the chamber through an additional door (opening) at the top of the chamber, without affecting the measurement conditions.

During the measurement of self-noise and pressure drop, the tested silencer is connected to a centrifugal fan via three absorption silencers. The fan, noise source, and tested silencer are positioned outside the reverberation chamber, whereas the outlet is situated inside the chamber.

A procedure for determination of the SPL is measurement in points evenly distributed on the circular measurement path of 10 m in circumference by using microphone rotating boom. SPLs are measured in nine discrete microphone positions with an integration time of 30 s for each microphone position and are measured in 1/3 octave bands 100 Hz to 10 000 Hz. The B&K microphone, model no. 4146 was used and data were collected using a two-channels B&K analyser 2144. Microphone was calibrated before commencing the acoustic test. Background noise is recorded for each measurement series with airflow switched off, enabling the calculation of background correction K_1. The reverberation time is measured for four omnidirectional loudspeaker positions with three microphone settings. All sound power level calculations are completed using a dedicated calculation sheet. Due to the high signal-to-noise ratio (the difference between the signal and the background exceeds 15 dB) and excellent spatial uniformity (variance between signals from 12 positions of microphones <0.3 dB), the expanded uncertainty of the sound power level measurement in a reverberation room was estimated at the laboratory as $U = \pm 0.7$ dB. This reflects a high-precision class 1 laboratory measurement where environmental and instrumental errors are effectively minimised.

The flow velocity was determined using the so-called arithmetical calculation method described in the norm PN-ISO 5221:1994 'Distribution and division of air – Measurement procedures for airflow in the duct'. The mean velocity in the channel was determined using the log-Chebyshev method. Acoustic measurements were taken at four flow velocities, $\nu_1 = 3$ m/s, $\nu_2 = 6$ m/s, $\nu_3 = 9$ m/s, and $\nu_4 = 12$ m/s. Ambient pressure was measured by means of the Prandtl probe with a pressure difference converter. While for the duration of the flow noise measurement, the Prandtl probe was removed from the channel measurement space so as not to disturb the measured acoustic signal. Also the static pressure and temperature in the ducts were measured. The use of an automated robotic positioning system in flow measurement significantly reduced the uncertainty associated with probe orientation and spatial sampling. For a velocity of 12 m/s, the expanded uncertainty was estimated at 1.3 %, primarily dominated by the sensitivity of the differential pressure transducer at low dynamic pressures. Whereas, by using a high-precision differential pressure transducer with an absolute error of ±0.1 Pa, the expanded uncertainty at the lower velocity of 3 m/s was 2.86 %. To sum up, the uncertainty of the flow measurements did not exceed 3 %. On a logarithmic scale (decibels), a 3 % error in velocity measurements translates to a measurement uncertainty of approximately ±0.8 dBA. This value is close to the human ear's limit of discrimination (1 dB), meaning that the measurement uncertainty for airflow measurements in the duct under test is at an acceptable level for acoustic engineering.

4. Results

4.1. Flow noise

Table 2 and Table 3 present 1/3 octave band sound power levels and overall sound power level (L_W) values reported for the self-noise of all tested ducts. Table 2 contains results obtained for a single flow velocity, namely 3 m/s and 6 m/s. Table 3 presents the results for flow velocity appropriately 9 m/s and 12 m/s. The penultimate row of each tables contains the single-number values of sound power level (L_w). The last row of tables give the differences between the single-number sound power level studied steel duct and the aluminium duct (ΔL_w).

For each of the tested steel ducts, the self-noise is higher than for the aluminium duct. The lowest single-number sound power level values (L_w) were obtained for the aluminium duct at airflow velocities of 3 m/s, 6 m/s, and 12 m/s; only at 9 m/s was the L_w value, obtained for the aluminium duct, higher than for the tested steel ducts. Considering the estimated uncertainty of the duct's self-noise sound power level measurement of ±0.7 dB, it can also be concluded that at velocities of 3 m/s and 12 m/s, L_w values for the aluminium duct are the lowest compared to the other tested ducts. At velocities of 6 m/s, the difference between the aluminium and galvanised ducts, after taking into account the spread of measurement uncertainty, is 0.6 dB, favouring the galvanised duct. At 9 m/s, the sound power levels of the tested duct are comparable, differing by only 0.1 dB to 0.2 dB.

A positive difference in L_w as a single-number value between the tested steel ducts and the aluminium duct indicates how much higher the acoustic power of this duct is than the base – aluminium duct. The largest differences were observed at velocity of 12 m/s, up to 2.5 dB for the galvanised duct and 2.3 dB for the raw

TABLE 2. Sound power levels of self-noise for all studied duct at 3 m/s and 6 m/s airflow.

f_r [Hz]	Drawn aluminium	Raw steel sheet	Painted steel sheet	Galvanised steel sheet	Drawn aluminium	Raw steel sheet	Painted steel sheet	Galvanised steel sheet
	ν_1 = 3 m/s				ν_2 = 6 m/s			
100	25.2	27.3	27.1	25.9	31.9	32.9	32.8	31.6
125	26.4	28.8	28.9	27.5	33.1	34.5	34.6	33.1
160	28.2	30.6	30.1	28.6	34.9	36.3	35.8	34.3
200	28.3	30.4	30.2	30.0	35.1	36.0	35.9	35.6
250	31.6	36.7	34.5	32.3	38.3	42.4	40.2	37.9
315	34.0	35.0	34.6	33.6	40.7	40.7	40.3	39.3
400	33.6	35.4	35.2	34.5	40.3	41.1	40.9	40.2
500	32.7	34.1	33.7	33.1	39.4	39.8	39.4	38.8
630	32.7	34.1	33.5	32.9	39.4	39.8	39.2	38.6
800	30.5	32.0	31.9	30.8	37.2	37.7	37.5	36.5
1000	27.4	28.6	28.2	27.5	34.1	34.3	33.9	33.1
1250	25.3	26.8	26.5	25.6	32.1	32.5	32.2	31.3
1600	21.2	22.8	22.3	21.3	27.9	28.5	28.0	27.0
2000	18.2	19.7	19.5	18.4	24.9	25.4	25.2	24.1
2500	14.5	16.2	16.0	15.1	21.3	21.9	21.7	20.8
3150	13.0	15.1	14.6	14.4	19.7	20.7	20.2	20.1
4000	12.5	15.8	16.3	15.9	19.3	21.5	22.0	21.5
5000	13.0	15.9	16.1	15.5	19.7	21.5	21.7	21.2
6300	15.1	16.8	17.0	16.9	21.8	22.5	22.7	22.6
8000	10.0	11.5	11.5	11.6	16.7	17.2	17.2	17.3
10 000	8.7	11.2	10.9	10.9	15.4	16.9	16.5	16.6
$\boldsymbol{L_w}$	**41.6**	**43.7**	**43.0**	**42.0**	**48.3**	**49.4**	**48.7**	**47.7**
$\boldsymbol{\Delta L_w}$	–	**2.1**	**1.5**	**0.5**	–	**1.1**	**0.4**	−0.6

TABLE 3. Sound power levels of self-noise for all studied duct at 9 m/s and 12 m/s airflow.

f_r [Hz]	Drawn aluminium	Raw steel sheet	Painted steel sheet	Galvanised steel sheet	Drawn aluminium	Raw steel sheet	Painted steel sheet	Galvanised steel sheet
	ν_3 = 9 m/s				ν_4 = 12 m/s			
100	41.8	42.3	41.7	41.9	45.6	48.6	47.4	49.3
125	43.5	44.2	45.7	43.9	47.4	50.5	50.2	50.5
160	44.2	45.1	44.6	43.9	49.8	52.5	51.1	51.8
200	44.7	45.4	44.9	45.8	49.8	52.6	51.5	53.2
250	47.0	46.7	47.9	47.2	52.8	53.6	53.8	54.4
315	51.2	50.1	49.9	50.0	56.8	58.2	57.0	57.9
400	50.9	51.0	50.9	50.9	56.3	59.0	58.1	59.3
500	50.3	50.1	49.7	49.9	55.7	58.1	57.1	58.2
630	50.5	50.4	49.7	49.9	56.1	58.4	57.1	58.5
800	49.3	49.2	48.9	49.0	54.7	57.4	56.3	57.7
1000	47.0	46.4	45.9	46.2	52.5	54.5	53.3	54.7
1250	46.5	46.1	45.7	46.0	52.6	54.8	53.7	54.9
1600	43.5	43.2	42.6	42.9	50.5	53.1	51.8	53.3
2000	41.3	41.0	40.6	40.9	48.4	51.5	50.2	51.9
2500	38.2	37.9	37.6	37.9	45.6	48.8	47.6	49.2
3150	34.7	34.4	34.1	34.5	42.4	45.8	44.7	46.3
4000	32.4	32.3	32.2	32.5	40.0	43.8	42.8	44.3
5000	28.5	28.6	28.9	29.0	35.0	39.4	38.5	39.7
6300	27.6	28.1	27.9	28.9	32.5	36.5	35.8	36.9
8000	21.1	21.9	22.2	22.9	28.3	32.4	31.7	32.8
10 000	18.4	19.6	20.2	21.2	25.1	29.1	27.9	29.4
$\boldsymbol{L_w}$	**59.3**	**59.1**	**58.9**	**58.9**	**64.8**	**67.2**	**66.1**	**67.3**
$\boldsymbol{\Delta L_w}$	–	−0.2	−0.4	−0.3	–	**2.3**	**1.3**	**2.5**

steel duct. A 2.1 dB difference was also observed for the raw steel duct at a velocity of 3 m/s. The raw steel duct has the highest roughness among the tested ducts, which translates into L_w values. It can therefore be concluded that this type of duct tends to increase the sound power level in the flow. The influence of the roughness of the used duct on the generated noise is significant for ventilation systems, where the surface is a factor, due to higher flow velocities than for liquids and lower fluid viscosity. For the tested ducts, the range of the R_a parameter is from 0.80 µm to 12.50 µm (over a 15-fold difference). In ventilation technology, a duct with R_a = 0.80 µm behaves like a 'hydraulically smooth' duct, while a duct with R_a = 12.50 µm (raw sheet metal) enters the region of turbulent flow, where roughness directly increases the resistance coefficient λ. The 6 % uncertainty in the R_a measurement associated with the roughness measurement is negligible compared to the 1500 % difference between the sheet metal variants. Higher roughness increases the thickness of the boundary layer, which physically 'narrows' the duct lumen for core flow (Keli *et al.*, 2023). The wall roughness is therefore a source of so-called generated noise, because air 'catches' surface irregularities, creating vortices that are the source of sound. The uncertainty of the profilometer measurement does not change the fact that a duct with a high R_a will generate higher self-noise. In the case of galvanised duct, the galvanising method (hot-dip galvanisation, electroplating, or spraying) is important. In this study, galvanised sheet metal with a mill finish (as-rolled) was used, which refers to being hot-dipped in zinc after being hot-rolled and potentially pickled (the hot-dip method). This galvanising and finishing method results in a higher roughness suitable for spraying or painting as the next step of the metal process (Tatarek *et al.*, 2009). For the painted duct, which has lower roughness, the difference ΔL_w does not exceed 1.5 dB.

A velocity measurement uncertainty of 3 % translates to the SPL uncertainty of approximately ±0.8 dB, considering that acoustic power typically scales with v^6 (sound power is proportional to the sixth power of flow velocity). This uncertainty is within acceptable limits for most industrial acoustic assessments. Take into account the measurement uncertainty of surface roughness, the observed variations in sound parameters from flow characteristics across the ducts stem from genuine differences in a surface finish rather than instrumental error. The measurement uncertainty may marginally reduce the resolution of the comparison, but it does not undermine the fundamental findings of the study. The higher parameters of surface roughness had a more dominant effect on the sound power levels due to the increased turbulence intensity in the boundary layer. Literary studies suggested that roughness elements can amplify certain frequencies of sound due to the interaction between the turbulent boundary layer and the acoustic waves or lower them (Raposo *et al.*, 2021). The sound characteristics in ducts with uneven surfaces could lead to increased noise levels due to flow-induced vibrations, and the type and severity of roughness directly correlate with noise pollution within duct systems (Mori, Ishihara, 2020). The conclusion is that the roughness of ventilation system elements could alter sound radiation patterns and influence the overall acoustic performance of ducted systems. However, airflow within ducts must be understood to prevent undesirable sound generation and maintain efficiency. The study by Choy and Huang (2005) demonstrated that under certain conditions, airflow does not significantly alter acoustic performance up to a defined turbulence intensity, like in this case where the similar values of L_w are observed for studied ducts at 9 m/s. The conclusion idea is that the ventilation duct design should carefully balance aerodynamic and acoustic considerations.

4.2. Insertion loss

The insertion loss (IL or D) is often used to evaluate the acoustic attenuation performance of the object working in airflow conditions. The tested duct's acoustic performance is measured using insertion loss. The insertion loss D_i is defined as the reduction in sound power level measured downstream of the studied aluminium duct and after replacing it with the studied steel ducts. In the present study, a drawn aluminium duct, as a very smooth with minimal airflow friction was treated as a low-noise object ('silencer'). The insertion loss is calculated according to:

$$D_i = L_{W\mathrm{II}} - L_{W\mathrm{I}}, \tag{4}$$

where $L_{W\mathrm{I}}$ is the the sound power level in the considered frequency band, measured for aluminium duct ('silencer'), and $L_{W\mathrm{II}}$ is the sound power level in the same frequency band, measured for studied steel ducts.

This approach enables the evaluation of the acoustic effectiveness associated with the use of steel ducts of varying surface roughness. For the tested duct, the correlation between insertion loss and duct surface roughness was determined for fully turbulent flow. Each of the tested velocities results in fully turbulent flow, i.e., the Reynolds number for 3 m/s is 60 200, for 6 m/s – 120 397, for 9 m/s – 180 595, and for 12 m/s – 240 794. Table 4 and Table 5 present the calculated insertion loss values in 1/3 octave bands for the studied steel ducts. The insertion loss spectra in 1/3 octave bands for tested ducts are shown in Fig. 3. The cut-off frequency (f_{cutoff}), the lowest frequency at which a particular wave mode can propagate through the duct, was marked in Fig. 3. The vertical pink line indicates the cut-off frequency (f_{cutoff} = 214 Hz), meaning that below it frequency, in a 500 mm × 800 mm duct, waves propagate flat, while above it, transverse modes begin to occur.

Due to the use of the aluminium duct as a reference ('silent'), the following assumptions should be made when interpreting the insertion loss of the tested ducts:

- if the insertion loss represents positive values, then the aluminium channel is quieter than the tested steel channel which acts as a 'muffler,'
- if the insertion loss assumes negative values, then the tested steel ducts are quieter than the reference aluminium duct which acts as an 'anti-muffler.'

Insertion loss allows us to draw key conclusions regarding the influence of the physical structure of the channel walls on aeroacoustic phenomena. Generally, from Table 3 and Table 4 and Fig. 3, we can conclude that at low frequencies (<500 Hz), large fluctuations and high attenuation values (up to 3 dB) are visible. This is the region where channel geometry and eigenmode resonances have the greatest impact on acoustics. At medium and high frequencies (500 Hz to 5000 Hz), the graphs stabilise and oscillate around 0 dB to 1 dB. This indicates that the type of sheet material has less influence on sound level changes. At very high frequencies (above 5000 Hz) and at higher velocities (9 m/s and 12 m/s), a further increase in attenuation is observed (especially for painted sheet metal), which may be due to the absorption characteristics of the material or changes in the boundary layer structure at higher flow rates.

TABLE 4. Insertion loss values in 1/3 octave bands the studied steel duct (reference duct – aluminium duct) at 3 m/s and 6 m/s flow velocity.

f_r [Hz]	Raw steel sheet	Painted steel sheet	Galvanised steel sheet	Raw steel sheet	Painted steel sheet	Galvanised steel sheet
	ν_1 = 3 m/s			ν_2 = 6 m/s		
100	3.4	2.8	2.9	2.4	1.8	1.9
125	1.0	1.6	1.5	−0.1	0.6	0.5
160	2.0	0.7	0.7	1.0	−0.3	−0.4
200	1.7	1.7	1.7	0.7	0.6	0.7
250	1.1	1.7	1.4	0.0	0.7	0.3
315	1.5	1.5	2.5	0.4	0.5	1.4
400	2.0	0.9	1.1	0.9	−0.1	0.1
500	0.7	0.7	0.8	−0.4	−0.3	−0.2
630	0.8	0.5	1.1	−0.3	−0.5	0.1
800	0.6	0.7	0.5	−0.5	−0.4	−0.5
1000	1.0	0.5	0.5	−0.1	−0.6	−0.5
1250	1.3	0.6	1.3	0.3	−0.4	0.3
1600	1.9	1.1	1.3	0.8	0.1	0.2
2000	0.8	1.3	1.2	−0.2	0.3	0.2
2500	1.6	0.9	1.6	0.6	−0.2	0.6
3150	0.7	1.2	0.9	−0.3	0.1	−0.1
4000	0.2	1.2	0.7	−0.8	0.1	−0.4
5000	1.0	1.2	1.1	−0.1	0.1	0.1
6300	0.8	1.4	1.1	−0.2	0.4	0.1
8000	0.6	1.7	1.4	−0.5	0.7	0.4
10 000	−0.3	1.3	0.5	−1.3	0.3	−0.6

TABLE 5. Insertion loss values in 1/3 octave bands the studied steel duct (reference duct – aluminium duct) at 9 m/s and 12 m/s flow velocity.

f_r [Hz]	Raw steel sheet	Painted steel sheet	Galvanised steel sheet	Raw steel sheet	Painted steel sheet	Galvanised steel sheet
	ν_3 = 9 m/s			ν_4 = 12 m/s		
100	2.2	1.8	2.2	2.3	1.5	3.0
125	0.4	0.8	1.0	0.6	1.4	0.6
160	0.7	−0.6	0.0	−0.3	−1.4	−1.2
200	0.8	0.5	1.3	0.1	−0.1	0.6
250	0.4	0.9	1.2	−0.2	0.1	0.1
315	0.5	0.4	1.8	−0.4	−0.3	0.6
400	1.1	0.2	0.7	0.5	−0.4	−0.1
500	−0.2	−0.2	0.2	−0.2	−0.3	−0.5
630	−0.2	−0.4	0.6	−0.3	−0.6	−0.2
800	−0.2	−0.3	0.1	−0.4	−0.5	0.0
1000	0.1	−0.5	0.1	0.6	0.0	0.2
1250	0.4	−0.3	0.8	0.1	−0.6	0.3
1600	0.9	0.1	0.7	−0.1	−1.0	−0.2
2000	0.0	0.3	0.5	0.4	0.7	0.6
2500	0.8	−0.1	1.1	0.4	−0.2	0.5
3150	0.1	0.4	0.8	0.4	0.6	0.7
4000	−0.5	0.2	0.2	0.2	0.7	0.5
5000	−0.1	0.1	0.4	0.2	0.2	0.4
6300	−0.2	0.4	0.6	0.9	1.3	1.3
8000	−0.4	0.5	0.5	1.2	2.1	1.6
10 000	−1.1	0.5	0.0	1.1	2.6	1.8

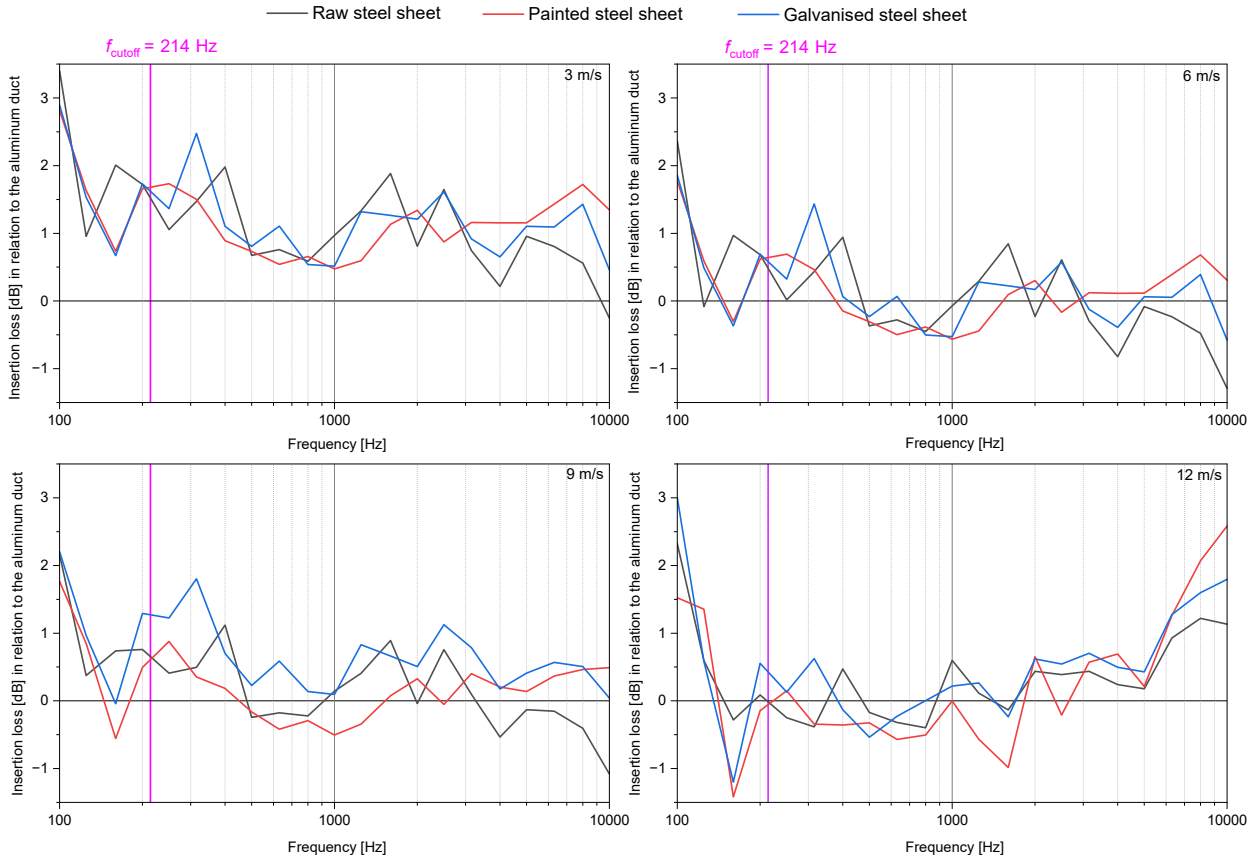

FIG. 3. Insertion loss spectra in 1/3-octave bands for studied steel ducts (aluminium duct as references) depending on the airflow velocity.

The effect of flow velocity on the differences in insertion loss between sheet metal types is noticeable but not dominant, particularly at flow velocities of 3 m/s and 6 m/s. Positive insertion loss values at 3 m/s indicate that the aluminium duct is an effective attenuator ('muffler') compared to the other tested ducts. For example, the highest insertion loss is 2 dB at 400 Hz compared to the raw duct, but compared to the galvanised duct, the highest insertion loss is 2.5 dB at 315 Hz. At 6 m/s, the aluminium duct attenuates sound in the ranges of 250 Hz to 500 Hz, 1000 Hz to 3000 Hz, and above 6000 Hz (only painted and galvanised duct). But, at 9 m/s, only the galvanised duct exhibits positive insertion loss, not exceeding the maximum value of 2 dB at 315 Hz. These results may indicate that the aluminium duct attenuates sound better than the galvanised duct at this airflow speed. At the highest velocity, deep minima (drops below 0 dB) are visible, suggesting that at this velocity, the airflow generates its own noise (self-noise), which may exceed the attenuation, or that surface resonances in the sheet metal are excited.

Figure 3 also shows that painted sheet metal often exhibits different minima than unpainted sheet metal. The paint layer changes the material's internal damping (damping), which shifts the resonance points. It is worth noting that, the painted channel has lower insertion loss values than the other two channels up to approximately 5000 Hz. It may mean that the painted duct will be comparable to the aluminium duct in terms of acoustic efficiency. It could also be connected with the similar roughness of these ducts. However, the insertion loss value increases above 5000 Hz for this duct, which means that the aluminium duct is acoustically better above this frequency than the painted one. A distinct resonance peak observed at 315 Hz for the galvanised steel sheet, which remains stationary across all flow velocities (3 m/s to 12 m/s), indicates a structural or geometric resonance rather than an aerodynamic phenomenon. This is likely caused by the specific panel resonance of the galvanised duct walls or the inherent damping properties of the zinc coating, which coincides with the excitation of the first transverse acoustic modes above the 214 Hz cut-off frequency.

In summary, it can be concluded that aluminium duct is acoustically more advantageous for use at lower airflow speed. However, at higher flow velocity, it may not be. This observation can be useful for designers because of the higher cost of aluminium ducts than steel ones. From single-number values of L_w studied, ducts can be observed that the roughness of the ducts affects their insertion loss. As airflow velocity increases, the differences between the materials become less pronounced, suggesting that aerodynamic effects dominate over material surface treatments at higher velocities.

4.3. Correlation between geometric surface properties and acoustic parameters

In this study, the Pearson linear correlation coefficient was applied to assess the strength and direction of the relationship between the channel surface roughness and the sound power level. The choice of this method is motivated by the continuous nature of the variables considered and by the assumption that, within the investigated range of operating conditions, the relationship between roughness and acoustic power can be approximated as linear. Moreover, the use of sound power level expressed in decibels contributes to the linearisation of the relationship and to variance stabilisation, which further supports the application of a linear correlation measure. Prior to the use of the Pearson correlation, the data were examined with respect to linearity, approximate normality of distributions, and the presence of outliers. These preliminary analyses indicated that the assumptions of the Pearson correlation were sufficiently satisfied for the considered dataset. Therefore, the Pearson coefficient was deemed an appropriate tool for quantifying the correlation between the geometric parameter (channel roughness) and the acoustic parameter (sound power level), providing an objective measure of the strength and direction of their linear relationship.

Table 6 summarises data on the roughness of the tested ducts and single-number values (and also in octave bands) of the sound power level depending on velocity. The last row of the table illustrates the Pearson coefficient.

The strongest relationship was observed at a flow velocity of 12 m/s (r = 0.91), where the correlation is almost perfectly linear. This indicates that, at this velocity, surface roughness becomes the dominant factor in noise generation: the boundary layer is sufficiently thin for roughness elements of about R_z = 50 µm to directly interact with the core flow, producing strong turbulence and regenerated noise. The difference between

TABLE 6. Pearson coefficient for studied ducts from the dependency between R_a and L_w (for single-number values and for octave band values) at different airflow speeds.

Air-flow speed [m/s]	Sound power level [dB]							
	L_w	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz	8000 Hz
12	0.91	0.90	0.81	0.84	0.86	0.86	0.82	0.85
9	−0.15	0.44	−0.69	−0.26	−0.48	−0.38	−0.29	0.55
6	0.32	0.39	0.59	0.46	0.38	0.44	0.81	0.55
3	0.78	0.38	0.39	0.38	0.36	0.37	0.36	0.36

aluminium (64.8 dB) and raw sheet metal (67.2 dB) is 2.4 dB, which is statistically highly significant. At 3 m/s (r = 0.78), the correlation is still high, suggesting that even at low flow rates, differences in surface structure are detectable by the measurement system, although absolute sound levels are much lower (around 42 dB). At 9 m/s, an almost zero correlation (r = −0.15) was obtained. Sound power levels are nearly identical for all materials (≈59 dB), which suggests that geometry-induced noise dominates and wall roughness temporarily loses its significance at this operating point. Parameters R_a and R_z are perfectly correlated in the dataset, but R_z = 50 µm is the physically meaningful factor, as it represents the actual height of obstacles interacting with the airflow. Considering the number of acoustical measurements (n = 12), low standard deviation, and very high correlation at 12 m/s, it may be concluded that the effect of roughness on acoustic power is statistically significant and not accidental.

Designers of ventilation systems often require sound power parameters in octave bands, because such data are directly necessary for designing installations in accordance with engineering practice and standard requirements. In connection with this, it is important to find the correlation between octave band results and surface roughness across the frequency spectrum. Table 6 shows also Pearson correlation coefficients for the octave bands of the tested ducts depending on their surface roughness at different air-flow speed. The octave band analysis at 12 m/s reveals a consistently high Pearson correlation ($r > 0.81$) across the entire frequency spectrum (125 Hz to 8000 Hz). The strongest correlation (r = 0.90) was observed at 125 Hz, indicating that surface roughness significantly enhances low-frequency turbulent pressure fluctuations and favours the formation of large-scale turbulent structures that generate low-frequency noise. It confirms that surface finish is a critical factor in flow-induced noise generation for this duct geometry. The spectral correlation analysis at 9 m/s demonstrates a significant shift in aeroacoustic behavior compared to higher velocities. While the 12 m/s data showed a dominant link between roughness and noise, at 9 m/s the Pearson coefficients fluctuate between weak positive and moderate negative values across most octave bands. This suggests that at this flow regime, surface-induced noise is no longer the primary sound source, and the small measured variations of L_w (often <0.3 dB) fall within the margin of experimental uncertainty. At this speed, the noise level generated by the wall roughness itself is so low that it is 'covered' by other sound sources, such as general turbulence in the channel or noise generated by the edges of the measuring holes. Consequently, the impact of surface morphology on the acoustic profile is effectively masked by broader flow turbulence at this specific velocity.

Spectral correlation analysis at 6 m/s reveals a consistent positive relationship ($r > 0$ across all bands) between the surface roughness (R_a) and sound power levels (L_w). Notably, a strong correlation (r = 0.81) is observed in the 4000 Hz octave band, indicating that at lower flow velocities, high-frequency noise components are the most sensitive to surface morphology. This suggests that roughness primarily generates high-frequency noise before becoming dominant across the spectrum at higher speeds. At the lowest tested velocity of 3 m/s, the spectral correlation analysis shows a consistent but weak positive relationship ($r \approx 0.37$) across all octave bands. The results are influenced by non-linearities, particularly a localised noise increase observed for the R_a = 1.6 µm variant, which suggests that at very low flow regimes, factors other than pure surface roughness – such as panel vibration or specific material damping – may play a more significant role.

The experimental data confirm that the influence of surface roughness on the sound power level (L_w) is non-linear and strictly dependent on the flow regime. At 12 m/s, a very strong linear correlation (r = 0.79) exists between roughness and noise. The thin boundary layer allows surface asperities to act as primary sources

of broadband noise, scaling with the sixth power of velocity. At 9 m/s, a 'dead zone' was identified where the correlation collapses ($r = -0.12$). In this state, general duct turbulence and geometric effects mask the acoustic signature of the surface texture. At the low-velocity regime (3 m/s to 6 m/s), roughness impacts the spectrum selectively, beginning with high-frequency bands (4000 Hz to 8000 Hz). At the lowest speeds, structural damping and resonances of the specific materials (e.g., the 315 Hz peak in galvanised steel) outweigh the influence of the R_a profile. The application of the Pearson correlation analysis across octave bands proved to be superior to single-number evaluations. It successfully pinpointed the exact frequency ranges where surface-induced noise is generated, providing a localised understanding of the aeroacoustic phenomena. This methodology could support a more precise approach to selecting insulation and attenuation components tailored to the specific roughness-induced noise profile of the ductwork.

5. Conclusions

This study investigates how the surface roughness of flat-oval ventilation ducts affects acoustic parameters, specifically focusing on self-noise and insertion loss. The three types of steel ducts – raw, galvanised, and painted – against a smooth-drawn aluminium duct used as a reference were compared. The experiments demonstrated that all tested steel ducts generated higher levels of self-noise than the smoother aluminium duct, with the most significant differences observed at the highest airflow velocity of 12 m/s, where the galvanised duct's noise level reached up to 2.5 dB higher. Painted steel ducts exhibited acoustic performance and self-noise levels very similar to the aluminium duct due to their comparably low surface roughness. A strong linear correlation was found between surface roughness and sound power level at 12 m/s, indicating that roughness becomes a dominant noise generator at high speeds. Conversely, at a medium velocity of 9 m/s, this correlation disappeared, suggesting that general turbulence masks the acoustic effects of surface texture at this specific speed.

The study confirmed that insertion loss in straight flat-oval steel ducts is generally very low, which is consistent with existing literature. At lower airflow velocities, such as 3 m/s and 6 m/s, the aluminium duct functioned as a more effective silencer than the steel ducts. However, as the airflow velocity increased, the differences in acoustic performance between the various materials became much less pronounced. This convergence indicates that at higher velocities, aerodynamic effects and turbulence dominate over the influence of the duct's material surface treatment. The spectral analysis further revealed that surface roughness impacts noise generation selectively, primarily affecting high frequencies between 4000 Hz and 8000 Hz at lower flow speeds.

A distinct structural resonance peak was observed in the galvanised steel duct at 315 Hz, which remained constant regardless of the airflow velocity. From a design perspective, while aluminium ducts are acoustically superior at low speeds, their higher cost may not be justified at high flow rates where aerodynamic noise prevails. Ultimately, predicting noise in HVAC systems remains difficult because the interaction between turbulent flow and surface roughness is non-linear and varies significantly across different flow regimes.

The main conclusion of this review study is that a comprehensive knowledge of turbulent flows over rough surfaces is still a long way off, even though there has been substantial development in this area. The main reason is that turbulent flows are erratic and unexpected, which makes accurate forecasting practically impossible. The large variety of roughness types, which significantly affect the flow dynamics in roughness sublayers, and the lack of comprehensive studies on the structure of turbulent flow are further problems.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contribution

Joanna Maria Kopania conceptualized the study and contributed to the analysis and interpretation of the data, and wrote the original draft. Kamil Wójciak performed the measurements, analysed, and interpreted the data. Patryk Gaj performed the measurements and contributed to the analysis. Grzegorz Bogusławski performed the measurements. All authors reviewed and approved the final manuscript.

References

1. Allen C.H. (1960), Noise control in ventilation systems, [in:] *Noise Reduction*, Beranek L.L. [Ed.], pp. 541–570, McGraw-Hill, New York.
2. American Society of Heating, Refrigerating and Air-Conditioning Engineers (2007), Sound and vibration control, [in:] *2007 ASHRAE Handbook: HVAC Applications*, Report, Chapter 47, ASHRAE Inc.
3. Bessac F., Guigou-Carter C., Lefebvre C., Bailhache S. (2018), Ductwork noise calculations: Main outputs of AcouReVe project, [in:] *39th AIVC Conference "Smart Ventilation for Buildings"*.
4. Boden H., Abom M. (1995), Modelling of fluid machines as sources of sound in duct and pipe systems, *Acta Acustica*, **3**: 545–560.
5. Boden H., Glav R. (2007), Exhaust and intake noise and acoustical design of mufflers and silencers, [in:] *Handbook of Noise and Vibration Control*, Crocker M.J. [Ed.], John Wiley & Sons, https://doi.org/10.1002/9780470209707.ch85.
6. Botejara-Antúnez M., González Domínguez J.G., García-Sanz-Calcedo J. (2023), Life cycle analysis methodology for heating, ventilation and air conditioning ductwork in healthcare buildings, *Indoor and Built Environment*, **32**(6): 1213–1230, https://doi.org/10.1177/1420326x231155146.
7. Choy Y.S., Huang L. (2005), Effect of flow on the drumlike silencer, *The Journal of the Acoustical Society of America*, **118**(5): 3077–3085, https://doi.org/10.1121/1.2047207.
8. Cummings A. (1980), Low frequency acoustic radiation from duct walls, *Journal of Sound and Vibration*, **71**(2): 201–226, https://doi.org/10.1016/0022-460X(80)90347-8.
9. Cummings A. (1983), Approximate asymptotic solutions for acoustic transmission through the walls of rectangular ducts, *Journal of Sound and Vibration*, **90**(2): 211–227, https://doi.org/10.1016/0022-460X(83)90529-1.
10. Cummings A. (2001), Sound transmission through duct walls, *Journal of Sound and Vibration*, **239**(4): 731–765, https://doi.org/10.1006/jsvi.2000.3226.
11. Cummings A., Chang I.-J., Astley R.J. (1984), Sound transmission at low frequencies through the walls of distorted circular ducts, *Journal of Sound and Vibration*, **97**(2): 261–286, https://doi.org/10.1016/0022-460X(84)90322-5.
12. Devenport W.J., Grissom D.L., Alexander W.N., Smith B.S., Glegg S.A.L. (2011), Measurements of roughness noise, *Journal of Sound and Vibration*, **330**(17): 4250–4273, https://doi.org/10.1016/j.jsv.2011.03.017.
13. Djeffal F. *et al.* (2021), Numerical investigation of thermal-flow characteristics in heat exchanger with various tube shapes, *Applied Sciences*, **11**(20): 9477, https://doi.org/10.3390/app11209477.
14. Fry A. (1988), *Noise control in building services*: *Sound Research Laboratories Ltd*, Pergamon Press, https://doi.org/10.1016/C2009-0-06822-6.
15. Henson P. (1986), *Computer programs incorporating the latest developments in calculation procedures for controlling ductborne noise in ventilation systems*, MSc. Thesis, South Bank University.
16. Herrin D.W., Seybert A.F. (2006), *Numerical methods for low-frequency HVAC noise applications*, ASHRAE report no. RP-1218.
17. Hersh A.S. (1983), Surface roughness generated flow noise, [in:] *AIAA 8th Aeroacoustics Conference*, https://doi.org/10.2514/6.1983-786.
18. Howe M.S. (1988), The turbulent boundary layer rough wall pressure spectrum at acoustic and subconvective wavenumbers, *Proceedings of the Royal Society A*, **415**(1848): 141–161, https://doi.org/10.1098/rspa.1988.0007.
19. Howe M.S. (1998), *Acoustics of Fluid-Structure Interactions*, Cambridge University Press.

20. International Organization for Standardization (2003), *Acoustics – Determination of sound power radiated into a duct by fans and other air-moving devices – In-duct method* (ISO Standard No. 5136:2003), https://www.iso.org/standard/28316.html.

21. Keli A., Rahnama S., Hultmark G., Hultmark M., Afshari A. (2023), Evaluation of elastic filament velocimetry (EFV) sensor in ventilation systems: An experimental study, *Sustainability*, **15**(3): 1955, https://doi.org/10.3390/su15031955.

22. Mori M., Ishihara K. (2020), Study on acoustic and flow-induced noise characteristics of L-shaped duct with a shallow cavity, *Noise Control Engineering Journal*, **68**(3): 209–225, https://doi.org/10.3397/1/376818.

23. Munjal M.L. (1987), *Acoustics of Ducts and Mufflers*, Wiley.

24. Pasanen P.O., Pasanen A.-L., Kalliokoski P. (1995), Hygienic aspects of processing oil residues in ventilation ducts, *Indoor Air*, **5**: 62–68, https://doi.org/10.1111/j.1600-0668.1995.t01-1-00010.x.

25. Raposo H., Mughal S., Bensalah A., Ashworth R. (2021), Acoustic-roughness receptivity in subsonic boundary-layer flows over aerofoils, *Journal of Fluid Mechanics*, **925**, https://doi.org/10.1017/jfm.2021.658.

26. Reynolds D.D., Bledsoe J.M. (1991), *Algorithms for HVAC Acoustics*, American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE).

27. Tahrour F., Djeffal F., C L., Benmachiche A.H. (2022), Numerical study to predict optimal configuration of wavy fin and tube heat exchanger with various tube shapes, *Journal of Renewable Energies*, **1**(1): 219–228, https://doi.org/10.54966/jreen.v1i1.1058.

28. Tatarek A., Kania H., Liberski P. (2009), Surface geometry of zinc coatings [in Polish: Geometria powierzchni powłok cynkowych], *Lakiernictwo*, **2**(58), https://www.lakiernictwo.net/dzial/142-aktualnosci-i-przeglad-rynku/artykuly/geometria-powierzchni-powlok-cynkowych,646/1.

29. VDI (2001), *VDI 2081: Noise reduction in air-conditioning systems*, Association of German Engineers, VDI, Düsseldorf.

30. Venkatesham B., Tiwari M., Munjal M.L. (2011), Prediction of breakout noise from a rectangular duct with compliant walls, *International Journal of Acoustics and Vibration*, **16**(4): 180–190.

31. Yu Y., Krynkin A., Horoshenkov K.V. (2024), The effect of 3D surface roughness on acoustic wave propagation in a cylindrical waveguide, *Wave Motion*, **128**: 103304, https://doi.org/10.1016/j.wavemoti.2024.103304.

Archives of Acoustics Vol. 51, No. 2, pp. 333–344 (2026), https://doi.org/10.24423/archacoust.2026.4282

Research Paper

Masculinized or Feminized? Discriminant Analysis of Postmenopausal Women's Voices

Maja Pietras[(1)], Łukasz Piotr Pawelec[(2)*], Monika Krzyżanowska[(1)], Anna Lipowicz[(2)]

[(1)] *Department of Human Biology, Faculty of Biological Sciences, University of Wroclaw* Wroclaw, Poland

[(2)] *Department of Anthropology, Wroclaw University of Environmental and Life Sciences* Wroclaw, Poland

*Corresponding Author: lukasz.pawelec@upwr.edu.pl

Received July 20, 2025; revised November 2, 2025; accepted January 26, 2026; available online January 30, 2026; version of record April 7, 2026; published issue June 24, 2026.

This study investigates the degree of vocal variation between men and pre- and postmenopausal women. The sample comprised 108 volunteers aged 18 to 66, divided into control and validation groups. Each participant was subjected to voice recordings of five sustained vowels. Acoustic parameters were extracted using Praat software. The most significant parameters in intergroup correlation between the canonical discriminant function and acoustic variables were: fundamental frequency (F0), shimmer, harmonics-to-noise ratio (HNR), and intensity. Premenopausal female voices were labeled with 97 % correctness and male voices with 95.5 % correctness. Interestingly, 65.5 % of postmenopausal women were accurately classified as female voices and on average they had lower vocal pitches compared to premenopausal women. The differences in male and female voices are probably due to the difference in the size of the larynx and the length of the vocal cords. Hormonal changes during menopause may affect, but not significantly, the morphology of the laryngeal structures which develop during childhood and adolescence.

Keywords: acoustical analysis, aging, fundamental frequency, hormones, menopause.

1. Introduction

1.1. Voice as a crucial biological trait

The source-filter theory is a fundamental concept in understanding the production of speech in mammals. It posits that the generation of speech involves two primary components: the source, which refers to the sound produced by the vocal cords in the larynx, and the filter, which represents the shaping of this sound by the supralaryngeal vocal tract (SVT) (Fitch, 2000; Taylor, Reby, 2010; Tokuda, 2021). Both source and filter characteristics are essential for effective communication. In many species, variation of voice characteristics contributes to individual distinctiveness and seems to be especially crucial for kin recognition, specifically for mothers and their offspring (Taylor, Reby, 2010). Primary factors that contribute to acoustic characteristics and quality of voice are: sex, age, body size/shape, and health which are mainly associated with physiological changes in sex hormone levels (Leongómez *et al.*, 2021; Puts, 2005). Age-related changes are connected to morphological changes in childhood, puberty and elderly. With age, muscle tone decreases, the elasticity and moisture content of the vocal fold tissues decline, and the length and thickness of the vocal folds change (Sataloff *et al.*, 2017).

1.2. Sexual dimorphism in voice parameters

Sexual dimorphism, the biological differences between men and women of the same species, extends to various aspects of human physiology, including voice. The role of sex hormones is crucial in shaping sexual dimorphism in voice parameters, with a focus on vocal pitch, timbre and voice quality. Pitch stands as one of the most prominent markers of sexual dimorphism in the human voice (Rosenfield *et al.*, 2020). Men typically have longer vocal cords and vocal tracts compared to women, leading to a lower pitch and narrower spacing of formant frequencies in men. While the evolutionary explanations for these sex differences are not fully understood, some evidence suggests a role for intrasexual competition. Men's pitch is approximately half as high as women's voices. This difference in pitch is largely attributed to men having vocal cords that are 60 % longer than those of women, a much larger difference compared to the 7 % disparity in height between both sexes (Puts *et al.*, 2007). The intersexual selection suggests correlations between female mate preferences and male voice characteristics, with women showing preferences for lower-pitched voices in potential mates. These preferences may reflect underlying genetic fitness or indicators of mate quality, such as physical size, health, and testosterone levels (Pisanski *et al.*, 2018b). Sex hormones significantly affect the vocal folds due to the presence of receptors on both androgen and estrogen hormones on them (Abitbol *et al.*, 1999; Aufdemorte *et al.*, 1983; Kirgezen *et al.*, 2017; Newman *et al.*, 2000; Voelter *et al.*, 2008). Studies indicate that androgens play a crucial role in development, structure and function of the human larynx. Specifically, androgens induce the hypertrophy of thyroarytenoid muscles, leading to a deepening of the voice pitch (Damrose, 2009; Huang *et al.*, 2015). During puberty, testosterone levels rise in men, inducing elongation and thickening of the vocal folds, which subsequently leads to a lower voice pitch. Conversely, women typically exhibit shorter and thinner vocal folds, resulting in a higher pitch. This difference in length can be attributed to the secondary descent of the larynx, a feature specific to men that occurs during puberty (Markova *et al.*, 2016). The impact of sex hormones on voice parameters extends beyond puberty, with hormonal fluctuations throughout the menstrual cycle and pregnancy exerting notable effects on vocal function in women (Pisanski *et al.*, 2018a). Variations in estrogen and progesterone levels during the menstrual cycle impact vocal fold tissue hydration and vascularization, leading to fluctuations in pitch and voice quality (Zamponi *et al.*, 2021).

1.3. Voice characteristics in menopausal women

Menopause is defined as the cessation of ovarian function and the decline in sex hormone levels, particularly estrogens, for at least 12 months (Lay *et al.*, 2020). Menopause and its symptoms affecting voice characteristics, is still a relatively new area of research. The change in hormone levels due to menopause can significantly affect vocal mechanisms, resulting in lower fundamental frequency and changes in voice quality, but the findings are inconclusive (Damrose, 2009; Huang *et al.*, 2015; Markova *et al.*, 2016). Some studies indicate that voice pitch is a key parameter affected by menopause-related hormonal changes. In meta-analysis, LÃ and Ardura (2022) stated that pitch is 0.94 semitones lower in post – as compared to premenopausal women. While notable, the extent of these declines falls below the threshold of perceptible difference and comfortably surpasses the threshold required to differentiate between female and male voices. During menopause, decreased estrogen levels may contribute to vocal fold atrophy, stiffness, dryness and throat clearing (Hamdan *et al.*, 2017; Shankar *et al.*, 2022). On the other hand, Hamdan *et al.* (2017) found no significant difference in the acoustic parameters between the pre- and postmenopausal groups. Some research suggests noticeable variations in acoustic parameters in both pre- and postmenopausal women (LÃ, Ardura, 2022) however, other studies show inconclusive results (Hamdan *et al.*, 2017). Both, male and female voices change over time – for example pitch (fundamental frequency, F0) in the case of men increase while in women – decrease (Tykalova *et al.*, 2021). It means that male and female vocal pitch becomes more similar with age. Thus, the main hypothesis of this study was to investigate whether postmenopausal women's voices are more similar to male or premenopausal female voices. The aim of the study was to apply discriminant function analysis for the classification of postmenopausal female voices to one of the groups: male or female voices and to establish the degree of method validity. What is more, it also identifies which acoustics parameters were fundamental for this assessment.

2. Methods

2.1. Participants

This study involved volunteers aged 18 to 65 years (mean age = 31.7 y., sd = 11.8 y.). The material consisted of two groups:

– first group: 44 men (mean age = 37.45 y., sd = 13.45 y., range: 20.5–66.9 y.) and 35 premenopausal women (mean age = 32.42 y., sd = 11.52 y., range: 18.2–50.6 y.),

– second group: 29 women in postmenopausal period (mean age = 57.18 y., sd = 4.51 y., range: 50.8–65 y.).

Research was conducted in accordance with the requirements of the declaration of Helsinki. The study was approved by the local ethical committee (Bioethics Committee at the Wroclaw Medical University, consent number: KB – 25/2021). All patients provided written consent before inclusion in the study.

2.2. Preliminary questionnaire

Each participant of the study was firstly asked to complete a preliminary questionnaire containing inclusion and exclusion criteria. These were questions about factors which may impact on voice quality, especially: head/neck medical history of trauma and treatments, malocclusions, hearing and speech defects, being ill on the day of examination, cigarette smoking, drinking alcohol on the day prior to the day of examination, voice-over work (i.e., working as a teacher, singer, sales representative, instructor etc.), COVID-19 disease history, use of hormonal agents (i.e., oral contraceptive for women). Moreover, women from the first group were asked about the current phase of the menstrual cycle. None of the participants answered affirmatively to any of the questions regarding the presence of the aforementioned inclusion factors. From the first group 25 women were in the menstrual phase, 27 in the follicular one, 43 in luteal one, and 9 of whom had ovulation. In the second group, for obvious reasons, the phase of the menstrual cycle on the day of examination was not defined. All of them declared that they were postmenopause.

2.3. Anthropometric data

Each participant had their height and weight measured. Body height was measured using an anthropometer with a range 0 cm to 200 cm and a precision to 0.1 cm. Body weight was measured using an InBody 270 electronic scale with an accuracy of 0.1 kg. Finally, body mass index was calculated using the following equation:

$$\mathrm{BMI} = \frac{\text{body mass}\,[\mathrm{kg}]}{\text{body height}\,[\mathrm{m}]^2}.$$

2.4. Voice recording procedure and acoustic analysis

The voices of the participants were recorded using the same equipment and equal acoustic conditions. The recording equipment consisted of a Shure SM 58 SE dynamic cardioid microphone with a frequency response 50 Hz to 15 kHz situated on a tripod, an IMG Stageline MPA-202 amplifier with 45 dB sound amplification and a low-frequency cutoff of 60 Hz and a Dell Latitude E6400 computer with an integrated sound card. The distance between the tip of the mouth and microphone and an angle between midline of the face and microphone were the same for each participant and were 15 cm and 0°, respectively. The recording conditions were also the same for all participants – the silent room (acoustic background measured with a digital sound level meter Benetech GM1351 ~39 dB), sitting position, acoustic cabin Mozos Mshield (microphone inside), the same time of the day (9 a.m. to 12 a.m.) and season of the year (autumn). Each participant was asked to speak aloud five vowels /ɑ:/, /ɛ:/, /i:/, /ɔ:/, /u:/ with sustained phonation lasting 3 s, with a 1-second break after each.

All sound files were recorded with the sampling frequency of 44.1 kHz and 16-bit resolution as uncompressed (.wav) mono files.

All data was subsequently analyzed with Praat software version 6.2.06 (Boersma, Weenink, 2019) using the middle fragment of each vowel of equal length (0.2 s) to determine acoustic parameters. Those were mainly F0,

formant frequencies (F1–Fn), and intensity. F0 is the perceived pitch of an individual voice, determined by the rate of vocal cord vibration and it varies among individuals based on factors such as age, sex, and health (Singh, 2019). Formant frequencies (formants) are resonant frequencies that shape the quality of vowel sounds in speech. They result from the acoustic filtering effects of the vocal tract on the sound produced by the larynx. Different vowels are characterized by distinct patterns of formant frequencies, which contributes to vowel identification (Pisanski *et al.*, 2016). Intensity refers to a pressure at which a sound is emitted, determining it as the loudness of the sound perceived by the listener (Zhang, 2016). Many studies prove that both pitch and timbre can indicate the individual body size and shape among adult men and women. Besides timbre and mean pitch, certain voice parameters may also hint at differences in height, weight and body circumferences, such as minimum F0, maximum F0, and F0 variability (Pisanski *et al.*, 2014; 2016; Pawelec *et al.*, 2022; Teixeira *et al.*, 2013). In addition to these primary parameters, other factors such as jitter, shimmer, and harmonics-to-noise ratio (HNR) contribute to the acoustic characteristics of the voice were computed. Jitter and shimmer are measures of the variations in F0 and intensity. Jitter refers to the cycle to cycle variations in F0, while shimmer quantifies the cycle to cycle variations in amplitude (Teixeira *et al.*, 2013). HNR is a measure used to quantify the balance between harmonic components and noise in the speech. It reflects the degree to which the sound consists of harmonically related components, which are characteristic of voiced sounds produced by the vocal folds, versus non-harmonic noise components, which may arise from various sources such as turbulent airflow or vocal fold irregularities (Murphy *et al.*, 2008). All acoustic parameters were averaged using 5 vowels' values for each participant.

2.5. Statistical methods

Basic descriptive statistics of physical and acoustics parameters were calculated (in the case of the first group separately for men and women) for both groups. To determine the discriminant functions of voice acoustics parameters for men and women the first group was used as a control group. The sex of participants was taken as an independent (grouping) variable, and nine voice parameters: F0, F1–F4, jitter, shimmer, HNR, and intensity as dependent variables. The second group containing data of postmenopausal women was a validation group. The linear discriminant analysis (LDA) method was used. To assess the differences of premenopausal women, postmenopausal women and men vocal pitch the one-way analysis of covariance (ANCOVA) including age and BMI as confounding variables and Tukey's HSD post-hoc test for unequal counts were applied. The Statistica 13.5 software (1984–2017 TIBCO Software Inc. Palo Alto, California, USA) was applied for all analyses. The significance level set to $p < 0.05$ was considered significant.

3. Results

3.1. Descriptive data

Descriptive statistics presenting central tendency and dispersion measures of the sample were shown in Table 1.

3.2. Linear discriminant analysis (LDA)

The discrimination of the participants' sex based on selected voice parameters was highly significant (Wilks' $\lambda = 0.17$, $F = 32.77$, $p < 0.001$). The significant acoustic characteristics for discriminant analysis were F0, shimmer, HNR, and intensity, therefore they were used for the next model. When these four variables were taking into account once again all of them remained significant (Wilks' $\lambda = 0.19$, $F = 81.39$, $p < 0.001$) thus these variables were used for all subsequent analyses (Table 2).

There was only one canonical discriminant function which was statistically significant ($x2 = 126.47$, $p < 0.001$, eigenvalue: 4.4) and its equation was as follow:

$$\mathrm{D1} = -0.57 - 0.05\ \mathrm{F0} + 19.4\ \mathrm{shimmer}\ + 0.12\ \mathrm{HNR} + 0.07\ \mathrm{intensity}. \tag{1}$$

Table 1. Descriptive data of a control and a validation group.

Trait	Control group (N = 79)								Validation group (N = 29)			
	Premenopausal women (n = 35)				Men (n = 44)				Postmenopausal women			
	Mean	Sd	Minimum	Maximum	Mean	Sd	Minimum	Maximum	Mean	Sd	Minimum	Maximum
Age [years]	32.42	11.52	18.20	50.60	37.45	13.45	20.50	66.90	57.18	4.51	50.80	65.00
Body height [cm]	167.09	5.01	155.00	175.00	179.91	5.91	162.00	190.00	162.94	6.96	146.00	176.90
Body mass [kg]	67.14	12.90	40.00	106.70	85.12	17.18	57.00	135.00	71.49	11.90	54.30	103.10
BMI [kg/m^2]	24.02	4.36	16.65	38.70	26.23	4.81	18.40	38.20	26.97	4.42	20.86	34.97
F0 [Hz]	204.57	23.73	143.68	273.72	120.74	22.84	90.60	174.63	183.74	22.98	148.18	246.06
Jitter [%]	0.42	0.24	0.11	1.22	0.44	0.33	0.16	2.30	0.35	0.15	0.14	0.75
Shimmer [%]	3.90	2.77	1.20	13.16	4.18	2.37	1.17	11.84	3.40	1.98	0.77	10.54
HNR [dB]	22.66	5.12	10.12	34.49	18.58	3.65	10.42	25.88	25.81	4.95	14.45	35.05
Intensity [dB]	72.99	8.05	55.88	89.21	76.54	8.97	61.95	90.26	73.36	15.66	61.86	149.05
F1 [Hz]	583.38	66.30	460.07	761.38	629.22	182.19	410.72	1113.27	575.09	62.02	469.42	777.23
F2 [Hz]	1570.71	223.54	1306.44	2636.22	1745.41	440.11	1283.67	3090.85	1538.51	205.95	1074.73	2302.56
F3 [Hz]	2883.25	280.79	2445.71	4171.51	2953.86	422.46	2530.88	4274.47	2879.53	219.68	2559.02	3728.81
F4 [Hz]	3990.10	351.97	3445.89	5706.09	4050.14	828.34	3387.96	6342.94	3945.70	395.36	3468.47	5736.74

Table 2. Voice characteristics and their meaning in discriminant function analysis.

Acoustic parameter	Wilks' λ	Partial λ	F	p
1st model				
F0	**0.64**	**0.27**	**189.33**	**<0.001**
F1	0.17	1.00	0.17	0.685
F2	0.17	1.00	0.08	0.781
F3	0.18	0.95	3.41	0.069
F4	0.18	0.95	3.62	0.061
Jitter	0.17	1.00	0.20	0.660
Shimmer	**0.19**	**0.92**	**5.68**	**<0.05**
HNR	**0.18**	**0.94**	**4.38**	**<0.05**
Intensity	**0.21**	**0.83**	**14.28**	**<0.001**
Final model				
F0	**0.73**	**0.25**	**217.35**	**<0.001**
Shimmer	**0.20**	**0.94**	**5.14**	**<0.05**
HNR	**0.20**	**0.92**	**6.03**	**<0.05**
Intensity	**0.23**	**0.81**	**17.40**	**<0.001**

The means of canonical discriminant function for men and women from a control group were presented in Table 3.

Table 3. Means of canonical discriminant function for both sexes – the control group.

Sex	Mean canonical discriminant function value
Men	1.85
Premenopausal women	–2.32

The highest significant intergroup correlation between the canonical discriminant function and acoustic variables was found for the F0 (Table 4).

Table 4. Intergroup correlations between canonical discriminant function and acoustics parameters.

Acoustic parameter	Canonical discriminant function
F0	–0.87
Shimmer	0.03
HNR	–0.22
Intensity	0.1

The classification matrix of training data (control group) was presented in Table 5. A priori classification probability for men was approximately 56 % and in the case of women 44 %. Women's voices were classified with a higher correctness (99 %) than men's (93.8 %). Only one woman was classified as a man while five men were classified as women. Ten postmenopausal women were classified as men and 19 as women based on voice signal. The classification correctness was 65.5 %. It means that acoustics parameters of postmenopausal women were more similar to premenopausal women than men's pitch, however the accuracy of classification was not 100 % (Table 5).

Table 5. Classification matrix of men and women according to discriminant function.

		Correct classifications [%]	Assessed as a man $p^* = 0.55696$	Assessed as a woman $p = 0.44304$
Training group (control)	Men	95.5	42	2
	Premenopausal women	97.1	1	34
Validation group	Postmenopausal women	65.5	10	19
	Σ	88.0	53	55

* *a priori* classification probability.

Those three groups differed from each other in terms of age and BMI as confounding variables significantly in pitch ($F = 114.16$, $p < 0.001$) and HNR ($F = 11.55$, $p < 0.001$). Mean pitch of postmenopausal women was lower than that of premenopausal women ($p = 0.0038$) but still significantly higher than men's ($p < 0.001$; Fig. 1). HNR of postmenopausal women was higher than that of premenopausal women ($p = 0.0072$) as well as that of men ($p < 0.001$; Fig. 2).

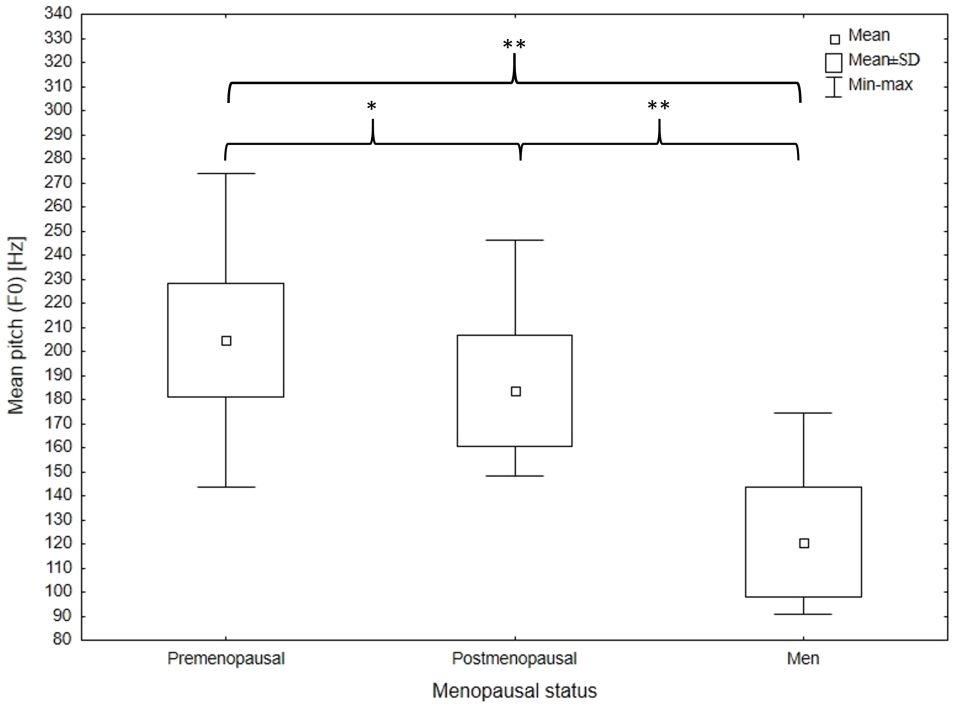

Fig. 1. Tukey's HSD post-hoc test for mean pitch (F0) differences between postmenopausal women and premenopausal women and men; $^*p < 0.01$, $^{**}p < 0.001$.

The F0 of postmenopausal women's voices was closer to that of premenopausal women, mostly separated from male voices. What is more, there were observed negative tendency of F0 and age for premenopausal female voices and positive tendency for male voices. Moreover, a negative trend was observed in postmenopausal women, as in premenopausal ones, but even stronger (Fig. 3).

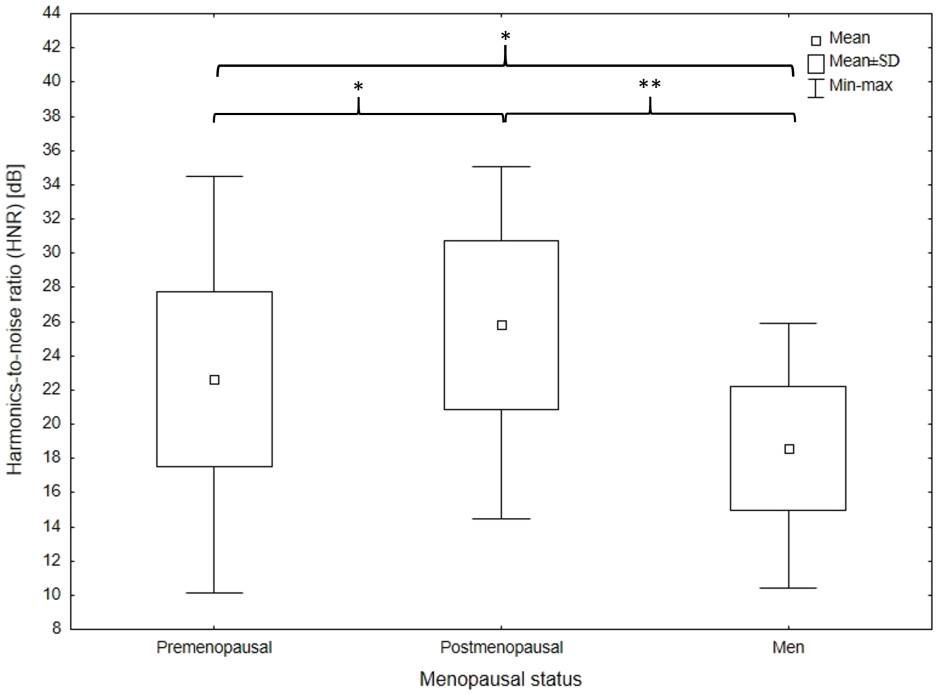

FIG. 2. Tukey's HSD post-hoc test for HNR differences between postmenopausal women and premenopausal women and men; $^{*}p < 0.01$, $^{**}p < 0.001$.

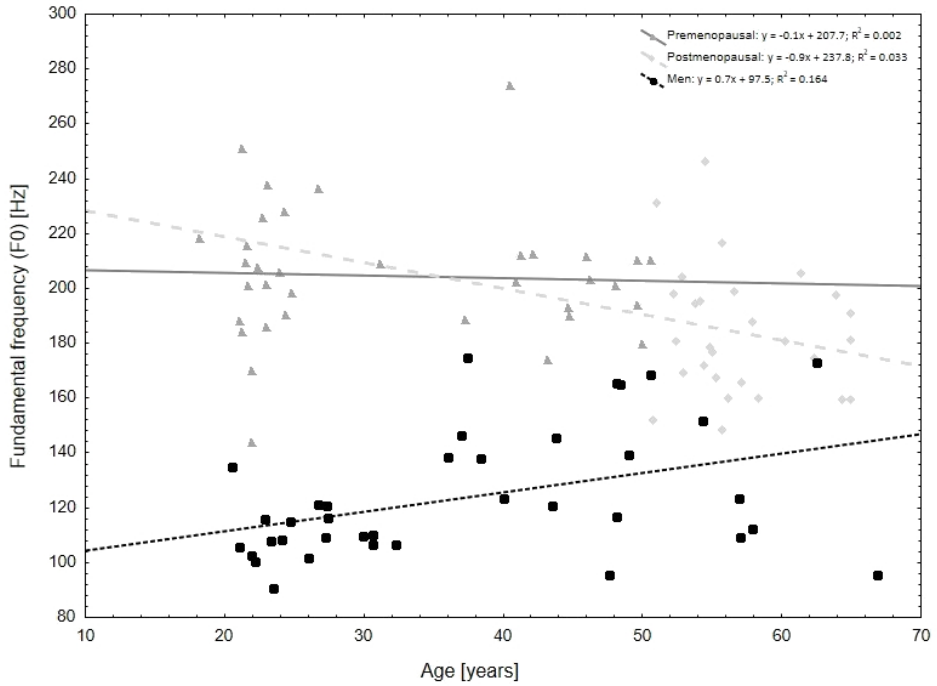

FIG. 3. Linear fitting of age and mean pitch (F0) for postmenopausal women, premenopausal women and men. Linear regression equations and R-square for each group.

4. DISCUSSION

This study was undertaken to apply discriminant analysis for the classification of postmenopausal females voices to one of the groups: males or females voices and to establish the degree of method validity. It appeared that most voices which belonged to postmenopausal women were classified correctly based on discriminant function to women's group. The mean F0 of postmenopausal women's voices was significantly lower than the mean F0 of premenopausal women's voices and higher than men's. Studies to date have shown consistent results: for men

their pitch increases by up to 35 Hz and for menopausal women their pitch decreases by 10 Hz to 35 Hz (Singh, 2019). Awan (2006) in a study of middle-aged women (40 to 59 years), who partly correspond to a group of postmenopausal women from this study, applying the discriminant analysis revealed that they were classified with 80 % correctness to the middle-aged group. The most significant discriminators of classification were vital capacity (VC) and the fundamental frequency standard deviation (called pitch sigma). The meta-analysis considered papers focusing on changes in vocal parameters after the menopause found that most studies had revealed after-menopause changes in the speaking fundamental frequency, SFF (n = 8) or F0 of sustained vowels (n = 10). The weighted average absolute difference for SFF was 10.10 Hz and for F0 based on sustained vowel /a/ 13.41 Hz (Lã, Ardura, 2022). This finding confirms the current study results, in which it was found that discriminant analysis function was also built based on the F0 of voice, among other acoustic parameters. The result of the current study is confirmed by the meaningful difference in pitch among male and female voices. The reasons for these discrepancies are the larynx size (~20 % greater in men) and vocal cords' membranous length (~60 % longer in men) which impact the voice properties of both sexes (Titze, 1989). The longer vocal folds, the lower F0 in men and women (Hollien, Moore, 1960). The voice differences revealed in this study come from the childhood and adolescence periods (first 20 years of life) when the larynx alongside vocal folds develop with various growth rates in men and women (Hirano, 1981). Those facts may support the hypothesis that, despite the altered hormonal profile in women during perimenopause, the anatomical differences in laryngeal structure created during progressive ontogeny are so strong that their voices are still categorized as female registers. It means that sex is a strong predictor of human voice pitch. Some evidence which enhances this statement is a study that examines event-related potentials (ERPs) of the brain as an answer to male/female voices. This findings revealed that:

– participants were able to correctly discriminate a sex of the adult speaker based on voice solely with 95 % correctness,
– the fastest brain responses were noticed for low-pitched voices categorized as a man and for high-pitch categorized as a woman (Hirano, 1981).

The authors stated that 'These results showed that a person's gender is in part derived from fundamental frequency (pitch) (...)' (Latinus, Taylor, 2012, p. 200). On the other hand, there is known that changes in voice acoustic parameters, such as decrease of mean F0 and an increase in shimmer, HNR, and voice turbulence index in female patient undergoing gender reassignment using testosterone, especially between the 3rd and the 4th month of therapy (Damrose, 2009). Moreover, another study indicated a significant difference in habitual pitch (HP) between menopausal women of comparable BMI who were on hormonal treatment (HT) or were not on HT. The higher value of HP was found in women on HT (Hamdan *et al.*, 2018). The authors explained those changes in voice quality by proliferative and hypertrophic estrogens' effect on vocal folds mucosa with an increase in mucus secretion and an antiproliferative progesterone's effect and decrease in glandular activity (Caruso *et al.*, 2000; D'haeseleer *et al.*, 2012). D'haeseleer *et al.* (2012) also found that in the case of postmenopausal women on HT had a significantly higher value of a speaking fundamental frequency (SFF) compared to those who were not on HT – the mean difference was approximately 14.2 Hz (D'haeseleer *et al.*, 2012). In another study the same authors observed a significant positive correlation between BMI and pitch in postmenopausal women who were not undergoing hormone therapy. Conversely, no correlation was found in either the premenopausal group or the postmenopausal group receiving hormone therapy. The association between BMI and pitch in postmenopausal women not on hormone therapy suggests a potential link to increased estrogen production in adipose tissue among individuals with elevated BMI (D'haeseleer *et al.*, 2011). These findings show that the sex hormones impact on voice parameters is apparent, but even without a menopausal hormonal therapy (MHT) the F0 is higher than the average for men based on (Teixeira *et al.*, 2013).

4.1. Limitations and future directions

This study has some limitations. First of all, the accurate determination of menopausal status of women is unknown as these women declared whether they were before or during/after menopause but the sex hormones'

level in their blood or saliva was not determined. In the future research it would be essential to precisely define the menopausal status of participating women (premenopause/perimenopause/postmenopause). The second limitation is the lack of postmenopausal women using MHT in our sample. According to other studies it seems to be important to compare both groups (MHT vs. not MHT) in each menopausal status (Caruso *et al.*, 2000; Hamdan *et al.*, 2018; Latinus, Taylor, 2012). The third limitation is the fact that non vocal apparatus imaging methods such as videolaryngoscopy, magnetic resonance imaging (MRI), or computed tomography (CT) were not performed. This step could help future researchers describe the quality of anatomical structures which take part in the speech production process (i.e., vocal folds and laryngeal cartilages, pharyngeal walls, soft palate, etc.). It would show what are the differences between these structures in men and in pre- and postmenopausal women. Another limitation of this study is the age of postmenopausal women (52 to 65 years), which means that these women have recently gone through the menopause. One would expect that voices of women who would be older than those from our sample might be more similar to male voices. It would probably decrease the accuracy of discriminant analysis and as a result more difficulties with sex assessment. Finally, discriminant function analysis was the only method applied to differentiate voices. In the further consideration it would be helpful to use some subjective methods, e.g., ask independent 'judges' to try to guess to whom – men, pre-, postmenopause women – belongs each voice.

5. Conclusions

The current study revealed that though the voice parameters of postmenopausal women were significantly different from those of premenopausal ones; postmenopausal women voices were still assigned with 65.5 % correctness to the women's group. The most discriminating voice parameter was F0, shimmer, HNR, and intensity. Controlling both, age and BMI, pitch and HNR differed between three studied groups. The average postmenopausal female voice was significantly lower than that of a premenopausal woman, but still higher than that of a man. These significant differences in vocal pitch between above-mentioned groups are probably due to the anatomical variation between men's and women's vocal apparatus structures that originated in childhood and adolescence (Hollien, Moore, 1960). The influence of sex hormones on a voice signal in those groups seems to be weaker but was not examined in the present study. Further research is needed to better understand the background of this phenomenon.

Fundings

The material for this study was collected thanks to funding under the project N020 – Fundusz Wsparcia Badań Naukowych – Bon Doktoranta Szkoły Doktorskiej, number N020/0008/20 granted by Wroclaw University of Environmental and Life Sciences, Poland.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions

Maja Pietras provided portions of the manuscript. Łukasz P. Pawelec contributed to the concept and design of the study, acquisition of the data, as well as statistical analysis, interpretation of the results and writing of the manuscript. Monika Krzyżanowska contributed to the concept and design of the study, provided key edits, and revised the manuscript. Anna Lipowicz contributed to the concept and design of the study, provided key edits, and revised the manuscript. All authors reviewed and approved the final manuscript.

Ethical approval

The study was approved by the local ethical committee (Bioethics Committee at the Wroclaw Medical University, consent number: KB – 25/2021). All patients provided written consent prior to inclusion.

Data availability statement

There are no linked research data sets for this paper. The database for this study is available on request.

Acknowledgments

We would like to thank The Jerzy Kukuczka Academy of Physical Education in Katowice for the opportunity to collect research material. We would also like to thank all participants who took part in this study.

References

1. Abitbol J., Abitbol P., Abitbol B. (1999), Sex hormones and the female voice, *Journal of Voice*, **13**(3): 424–446, https://doi.org/10.1016/S0892-1997(99)80048-4.

2. Aufdemorte T.B., Sheridan P.J., Holt G.R. (1983), Autoradiographic evidence of sex steroid receptors in laryngeal tissues of the baboon (papio cynocephalus), *The Laryngoscope*, **93**(12): 1607–1611, https://doi.org/10.1288/00005537-198312000-00013.

3. Awan S.N. (2006), The aging female voice: Acoustic and respiratory data, *Clinical Linguistics & Phonetics*, **20**(2–3): 171–180, https://doi.org/10.1080/02699200400026918.

4. Boersma P., Weenink D. (2019), Praat: Doing phonetics by computer, Computer program, Version 6.2.06, https://www.praat.org.

5. Caruso S., Roccasalva L., Sapienza G., Zappalá M., Nuciforo G., Biondi S. (2000), Laryngeal cytological aspects in women with surgically induced menopause who were treated with transdermal estrogen replacement therapy, *Fertility and Sterility*, **74**(6): 1073–1079, https://doi.org/10.1016/S0015-0282(00)01582-X.

6. Damrose E.J. (2009), Quantifying the impact of androgen therapy on the female larynx, *Auris Nasus Larynx*, **36**(1): 110–112, https://doi.org/10.1016/j.anl.2008.03.002.

7. D'haeseleer E., Depypere H., Claeys S., Baudonck N., Van Lierde K. (2012), The impact of hormone therapy on vocal quality in postmenopausal women, *Journal of Voice*, **26**(5): 671.e1–671.e7, https://doi.org/10.1016/j.jvoice.2011.11.011.

8. D'haeseleer E., Depypere H., Claeys S., Van Lierde K.M. (2011), The relation between body mass index and speaking fundamental frequency in premenopausal and postmenopausal women, *Menopause*, **18**(7): 754–758, https://doi.org/10.1097/gme.0b013e31820612d5.

9. Fitch W.T. (2000), The evolution of speech: A comparative review, *Trends in Cognitive Sciences*, **4**(7): 258–267, https://doi.org/10.1016/S1364-6613(00)01494-7.

10. Hamdan A.-L. *et al.* (2017), Vocal symptoms and acoustic findings in menopausal women in comparison to premenopause women with body mass index as a confounding variable, *Journal of Menopausal Medicine*, **23**(2): 117–123, https://doi.org/10.6118/jmm.2017.23.2.117.

11. Hamdan A.-L., Tabet G., Fakhri G., Sarieddine D., Btaiche R., Seoud M. (2018), Effect of hormonal replacement therapy on voice, *Journal of Voice*, **32**(1): 116–121, https://doi.org/10.1016/j.jvoice.2017.02.019.

12. Hirano M. (1981), The structure of the vocal folds, [in:] *Vocal Fold Physiology*, pp. 33–41, College-Hill Press.

13. Hollien H., Moore G.P. (1960), Measurements of the vocal folds during changes in pitch, *Journal of Speech and Hearing Research*, **3**(2): 157–165, https://doi.org/10.1044/jshr.0302.157.

14. Huang G., Pencina K.M., Coady J.A., Beleva Y.M., Bhasin S., Basaria S. (2015), Functional voice testing detects early changes in vocal pitch in women during testosterone administration, *The Journal of Clinical Endocrinology & Metabolism*, **100**(6): 2254–2260, https://doi.org/10.1210/jc.2015-1669.

15. Kirgezen T., Sunter A.V., Yigit O., Huq G.E. (2017), Sex hormone receptor expression in the human vocal fold subunits, *Journal of Voice*, **31**(4): 476–482, https://doi.org/10.1016/j.jvoice.2016.11.005.

16. Lã F.M.B., Ardura D. (2022), What voice-related metrics change with menopause? A systematic review and meta-analysis study, *Journal of Voice*, **36**(3): 438.e1–438.e17, https://doi.org/10.1016/j.jvoice.2020.06.012.

17. Latinus M., Taylor M.J. (2012), Discriminating male and female voices: Differentiating pitch and gender, *Brain topography*, **25**: 194–204, https://doi.org/10.1007/s10548-011-0207-9.

18. Lay A.A.R., do Nascimento C.F., Horta B.L., Chiavegatto Filho A.D.P. (2020), Reproductive factors and age at natural menopause: A systematic review and meta-analysis, *Maturitas*, **131**: 57–64, https://doi.org/10.1016/j.maturitas.2019.10.012.

19. Leongómez J.D. *et al.* (2021), Voice modulation: From origin and mechanism to social impact, *Philosophical Transactions of the Royal Society B*, **376**(1840): 20200386, https://doi.org/10.1098/rstb.2020.0386.

20. Markova D. *et al.* (2016), Age- and sex-related variations in vocal-tract morphology and voice acoustics during adolescence, *Hormones and Behavior*, **81**: 84–96, https://doi.org/10.1016/j.yhbeh.2016.03.001.

21. Murphy P.J., McGuigan K.G., Walsh M., Colreavy M. (2008), Investigation of a glottal related harmonics-to-noise ratio and spectral tilt as indicators of glottal noise in synthesized and human voice signals, *The Journal of the Acoustical Society of America*, **123**(3): 1642–1652, https://doi.org/10.1121/1.2832651.

22. Newman S.-R., Butler J., Hammond E.H., Gray S.D. (2000), Preliminary report on hormone receptors in the human vocal fold, *Journal of Voice*, **14**(1): 72–81, https://doi.org/10.1016/S0892-1997(00)80096-X.

23. Pawelec Ł.P., Graja K., Lipowicz A. (2022), Vocal indicators of size, shape and body composition in Polish men, *Journal of Voice*, **36**(6): 878.e9–878.e22, https://doi.org/10.1016/j.jvoice.2020.09.011.

24. Pisanski K., Bhardwaj K., Reby D. (2018a), Women's voice pitch lowers after pregnancy, *Evolution and Human Behavior*, **39**(4): 457–463, https://doi.org/10.1016/j.evolhumbehav.2018.04.002.

25. Pisanski K. *et al.* (2014), Vocal indicators of body size in men and women: A meta-analysis, *Animal Behaviour*, **95**: 89–99, https://doi.org/10.1016/j.anbehav.2014.06.011.

26. Pisanski K. *et al.* (2016), Voice parameters predict sex-specific body morphology in men and women, *Animal Behaviour*, **112**: 13–22, https://doi.org/10.1016/j.anbehav.2015.11.008.

27. Pisanski K., Oleszkiewicz A., Plachetka J., Gmiterek M., Reby D. (2018b), Voice pitch modulation in human mate choice, *Proceedings of the Royal Society B*, **285**(1893): 20181634, https://doi.org/10.1098/rspb.2018.1634.

28. Puts D.A. (2005), Mating context and menstrual phase affect women's preferences for male voice pitch, *Evolution and Human Behavior*, **26**(5): 388–397, https://doi.org/10.1016/j.evolhumbehav.2005.03.001.

29. Puts D.A., Hodges C.R., Cárdenas R.A., Gaulin S.J.C. (2007), Men's voices as dominance signals: Vocal fundamental and formant frequencies influence dominance attributions among men, *Evolution and Human Behavior*, **28**(5): 340–344, https://doi.org/10.1016/j.evolhumbehav.2007.05.002.

30. Rosenfield K.A., Sorokowska A., Sorokowski P., Puts D.A. (2020), Sexual selection for low male voice pitch among Amazonian forager-horticulturists, *Evolution and Human Behavior*, **41**(1): 3–11, https://doi.org/10.1016/j.evolhumbehav.2019.07.002.

31. Sataloff R.T., Kost K.M., Linville S.E. (2017), The effects of age on the voice, [in:] *Clinical Assessment of Voice*, 2nd ed., Sataloff R.T. [Ed.], pp. 221–240, Plural Publishing, San Diego, California.

32. Shankar R., Raj A., Rathore P.K., Meher R., Kaushik S., Batra V. (2022), Menopause and its effect on voice, *Indian Journal of Otolaryngology and Head & Neck Surgery*, **74**(Suppl 3): 5524–5530, https://doi.org/10.1007/s12070-021-02870-9.

33. Singh R. (2019), *Profiling Humans from their Voice*, Springer, Singapore.

34. Taylor A.M., Reby D. (2010), The contribution of source–filter theory to mammal vocal communication research, *Journal of Zoology*, **280**(3): 221–236, https://doi.org/10.1111/j.1469-7998.2009.00661.x.

35. Teixeira J.P., Oliveira C., Lopes C. (2013), Vocal acoustic analysis – Jitter, shimmer and HNR parameters, *Procedia Technology*, **9**: 1112–1122, https://doi.org/10.1016/j.protcy.2013.12.124.

36. Titze I.R. (1989), Physiologic and acoustic differences between male and female voices, *The Journal of the Acoustical Society of America*, **85**(4): 1699–1707, https://doi.org/10.1121/1.397959.

37. Tokuda I. (2021), The source–filter theory of speech, [in:] *Oxford Research Encyclopedia of Linguistics*, https://doi.org/10.1093/acrefore/9780199384655.013.894.

38. Tykalova T., Skrabal D., Boril T., Cmejla R., Volin J., Rusz J. (2021), Effect of ageing on acoustic characteristics of voice pitch and formants in Czech vowels, *Journal of Voice*, **35**(6): 931.e21–931.e33, https://doi.org/10.1016/j.jvoice.2020.02.022.

39. Voelter Ch. *et al.* (2008), Detection of hormone receptors in the human vocal fold, *European Archives of Oto-Rhino-Laryngology*, **265**: 1239–1244, https://doi.org/10.1007/s00405-008-0632-x.

40. Zamponi V., Mazzilli R., Mazzilli F., Fantini M. (2021), Effect of sex hormones on human voice physiology: From childhood to senescence, *Hormones*, **20**(4): 691–696, https://doi.org/10.1007/s42000-021-00298-y.

41. Zhang Z. (2016), Mechanics of human voice production and control, *The Journal of the Acoustical Society of America*, **140**(4): 2614–2635, https://doi.org/10.1121/1.4964509.

Archives of Acoustics Vol. 51, No. 2, pp. 345–358 (2026), https://doi.org/10.24423/archacoust.2026.4338

Research Paper

Indian Sign Language Alphabet Recognition and Speech Synthesis Using a Hybrid Deep Learning Approach

Aswani Sivan*, Chandra Eswaran

Department of Computer Science, Bharathiar University
Coimbatore, India

*Corresponding Author: aswanisivan44@gmail.com

Received September 17, 2025; revised January 9, 2026; accepted January 19, 2026; available online February 4, 2026; version of record May 5, 2026; published issue June 24, 2026.

Indian Sign Language (ISL) is vital for communication among India's hearing-impaired community. However, the lack of standardised datasets and reliable identification frameworks has hampered the use of ISL in modern assistive technology. This paper presents a deep learning-based solution to robust ISL alphabet identification, with an emphasis on both accuracy and practical use. A curated static ISL alphabet collection was created by combining authoritative visual references from the official Indian Sign Language website and the Ramakrishna Mission Vivekananda Educational and Research Institute (RKMVERI). Multiple deep learning models were trained and assessed, including CNN, ResNet-50, DenseNet-121, VGG16, MobileNetV2, and EfficientNet-B0, with a new hybrid CNN-ResNet architecture outperforming the others. 98 % classification accuracy is achieved by the suggested approach, outperforming individual baseline models. Furthermore, the framework is expanded to support real-time applications, combining webcam-based capture with immediate conversion of recognized signs to textual and synthesized vocal output. A comprehensive performance evaluation, including the confusion matrix analysis and ROC curves, demonstrates the solution's durability and practical applicability. This research enhances accessibility, promotes inclusive education, and prepares the path for scalable sign language translation systems in real-world human-machine interaction scenarios by enabling accurate and real-time ISL recognition with voice feedback.

Keywords: Indian Sign Language, deep learning, CNN–ResNet hybrid model, real-time recognition, text-to-speech, assistive technology.

Notations

Acc – accuracy metric,
C – set of ISL alphabet classes,
D – dataset of ISL alphabet images (curated from ISL website and RKMVERI),
$F1$ – $F1$-score metric,
P – precision metric,
R – recall metric,
ROC – receiver operating characteristic curve,
TTS – text-to-speech synthesis module,
X – input image frame (static or real-time),
Y – predicted output label (ISL alphabet),
Θ – trainable parameters of the neural network.

1. Introduction

Humans have an innate desire for communication, and sign language is a rich cultural and linguistic medium that allows the hearing-impaired population to connect through visual-spatial interactions (Mistry *et al.*, 2021).

Indian Sign Language (ISL), one of the variants widely used in India, with unique linguistic structures and cultural peculiarities. ISL has, however, fallen far behind more thoroughly studied sign languages, such as American Sign Language (ASL) (Qahtan *et al.*, 2023; Govindharajalu Kaliyaperumal, Gopalan, 2025), in terms of technology integration and the creation of assistive applications, despite its widespread use.

One of the most significant problems in increasing computational ISL identification is the paucity of big, standardised, and publicly available datasets. ISL datasets are small and dispersed and do not fully capture the diversity of the language, in contrast to ASL, which has substantial benchmark datasets (Gogoi *et al.*, 2025) to enable reliable model training and evaluation. This data scarcity impedes the construction of precise and generalizable machine learning models required for real-world ISL recognition.

The emergence of deep learning, particularly convolutional neural networks (CNNs) (Ashwanth *et al.*, 2023) and their advanced variations such as ResNet (Zhou *et al.*, 2021), DenseNet (He *et al.*, 2016), VGGNet (Huang *et al.*, 2017), MobileNet (Simonyan, Zisserman, 2015), and EfficientNet (Sandler *et al.*, 2018; Sharma, Singh, 2022) has resulted in notable advances in visual categorization and gesture recognition in recent years. These architectures have exhibited a greater capacity to extract discriminative spatial information from pictures, resulting in world-class performance in a variety of sign language identification challenges (Tan, Le, 2019). Nonetheless, there is a dearth of systematic comparative evaluations of these ISL-specific approaches, and there has been little research conducted on exploiting hybrid models to combine the characteristics of various network designs to improve ISL identification accuracy and robustness.

This research fills these gaps by presenting a deep learning-based system using a hybrid CNN–ResNet architecture for static ISL alphabet recognition. The Ramakrishna Mission Vivekananda Educational and Research Institute (RKMVERI, n.d.) ISL dictionary and the official ISL website were combined to create a carefully managed dataset that ensures linguistic and visual authenticity. The framework is expanded beyond static recognition to include a real-time recognition pipeline that combines webcam-based gesture capture with real-time translation into text and synthesized speech using a text-to-speech (TTS) engine (Kingma, Ba, 2015), enabling accessible communication for the community of people with hearing impairments.

This study makes several contributions:

1. Creating a standardised and curated ISL alphabet dataset.
2. A novel deep learning Architecture has been developing as a hybrid model.
3. Built a model of a real-time ISL-to-speech conversion system that goes beyond static gesture classification to enable interactive communication.
4. Extensive performance evaluation to highlight model dependability and practical feasibility utilizing measures like accuracy, recall, $F1$-score, confusion matrices, receiver operating characteristic (ROC) curves.

2. Related work

The manual feature extraction methods used in sign language recognition research have been superseded by deep learning-based frameworks. In early ASL and ISL investigations, techniques included motion trajectory analysis, skin-colour segmentation, and shape descriptors. These techniques have limitations, such as being susceptible to lighting and signer variability, but they performed well in controlled environments (Chollet, 2017; Nandi *et al.*, 2022). With the introduction of deep learning, convolutional neural networks (CNNs) emerged as the dominant approach for static sign detection. Research on ASL alphabets showed that CNNs trained on substantial benchmark datasets may get excellent accuracy levels, frequently above 95 % (Gupta *et al.*, 2025). Since then, architectures that provide gains in accuracy and computational efficiency for gesture categorization have been investigated, including VGG16 (Amangeldy *et al.*, 2022), ResNet (Pisharady, Saerbeck, 2015), DenseNet (Indian Sign Language Research and Training Center, n.d.), MobileNet, and EfficientNet.

The lack of consistent datasets has been the main reason for the slower progress in the ISL environment. In order to train CNN-based models with accuracies in the 85 % to 90 % range, a number of researchers tried to gather small-scale datasets for ISL alphabets and words (International Organization for Standardization, 1998; Kraśkiewicz *et al.*, 2024). It has also done the transfer learning (Rastgoo *et al.*, 2021) from ASL

datasets to ISL recognition; models like DenseNet have shown encouraging results, while dataset mismatch is still a problem (Karamanli, Aydogdu, 2019). With varying degrees of effectiveness, hybrid models like CNN–RNN architectures have been used for dynamic ISL gestures (Houtsma, 2007). Most of these solutions lacked real-time implementations, which are important for real-world applications.

Table 1 provides a comparative overview of a few chosen ASL and ISL investigations, highlighting the datasets, approaches, performance indicators, and constraints. The chart shows that although ASL research has access to extensive curated datasets and reliable benchmarks, ISL research still faces challenges such as a lack of real-time integration, voice output characteristics, and dataset scarcity (Koller, 2020).

Table 1. Comparative overview of prior ASL and ISL recognition studies.

Dataset source	Method	Reported accuracy [%]	Limitation
ASL hand shape dataset	Shape descriptors + HMM	82	Sensitive to lighting, signer-dependent
ASL fingerspelling benchmark	CNN	94	Restricted to static alphabets
ASL dataset	VGG16	95	High computational cost
ASL dataset	ResNet-50	96	Limited to static gestures
ASL dataset	DenseNet	95	Cross-dataset generalization weak
ASL dataset	MobileNet	92	Lightweight but less accurate
ASL dataset	EfficientNet	94	Requires large-scale training
Self-compiled ISL dataset	CNN	88	Small dataset, no benchmarking
ISL dataset (alphabets)	Transfer learning (VGG16)	89	Overfitting due to limited data
ISL dataset + ASL transfer	DenseNet	93	Dataset mismatch issues
ISL dynamic signs	CNN–RNN hybrid	91	No real-time system, no TTS

Highlighting dataset sources, methodologies, stated accuracy, and limitations, this table provides an overview of representative ASL and ISL recognition research. ASL research benefits from benchmark datasets and advanced architectures, whereas ISL studies remain constrained by dataset size, lack of real-time systems, and absence of speech integration.

3. Methodology

This section describes the entire process of creating an accurate alphabet recognition system for ISL, augmented with a pipeline for real-time speech synthesis. Data collection and preprocessing, model architecture design, training procedures, and the deployment of the real-time recognition and TTS framework are all included in the technique (Saini *et al.*, 2023; Pandey *et al.*, 2025).

The following steps comprise the ISL recognition methodology:

- categorization of ISL alphabets from static images using deep learning models,
- ISL recognition and TTS conversion in real-time with a CNN–ResNet hybrid model,
- the whole process is shown in Fig. 1, starting with the development and preprocessing of the dataset, then moving on to the training and assessment of the model, and concluding with a framework for speech synthesis and real-time recognition (Gogoi *et al.*, 2025).

3.1. Dataset preparation

A high-quality dataset (D) of static ISL alphabet images was curated by aggregating data from two authoritative sources: standardised visual references for ISL alphabets are available on the official ISL website, and the RKMVERI ISL dictionary, a widely recognized academic repository of ISL signs.

Each data point in the dataset is formally denoted as a tuple x_i, y_i, where

$$D = \{(x_i, y_i)\}_{i=1}^{N}, \qquad x_i \in R^{h\times w\times c}, \qquad y_i \in \{1, ..., 26\}$$

corresponds to the ground truth label representing the ISL alphabets A–Z, indexed numerically. The total dataset contains N samples, sufficient for training and evaluation purposes while maintaining linguistic and cultural

accuracy by virtue of the source credibility; x_i denotes the i-th image of dimension $h \times w \times c$ (height, width, color channels), y_i is the corresponding class label, where 1 to 26 represent the ISL alphabets A–Z, N is the total number of image samples.

Training a model that picks up a mapping is the issue:

$$f_{\_}(\theta) : X \to C,$$

where X is the image input space, C is the 26-class set, and θ is the model's trainable parameters.

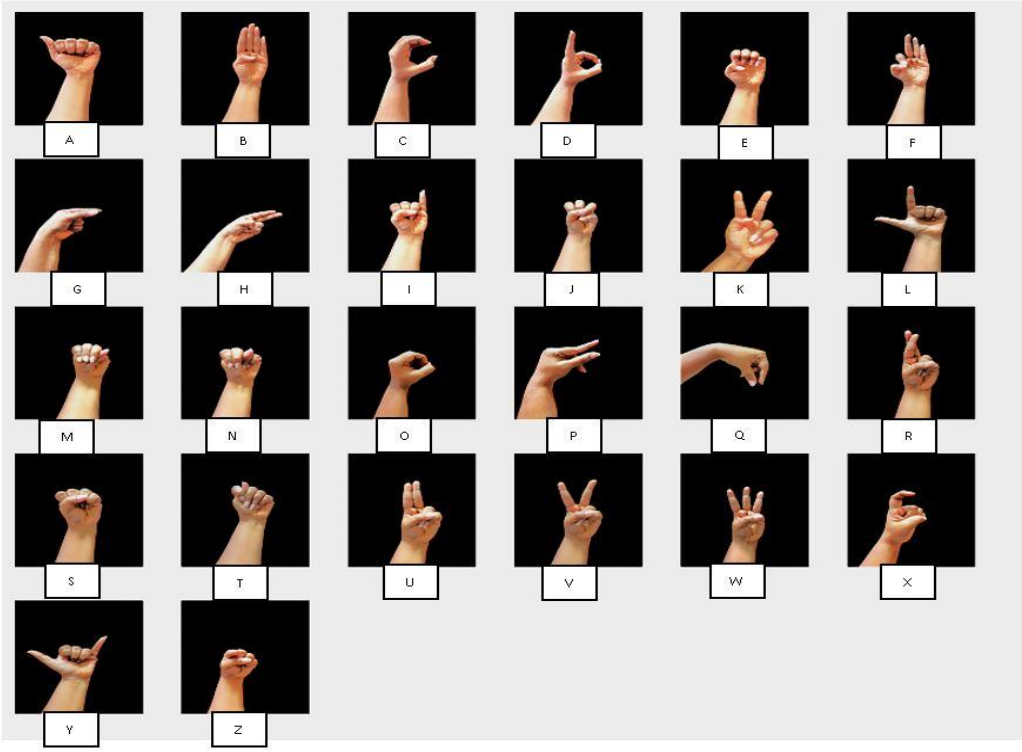

Fig. 1. Example images from the ISL dataset collected from the official ISL website and RKMVERI dictionary.

3.2. Preprocessing and augmentation

The following preprocessing procedures are applied to every image in order to support efficient training and model generalization:

- resizing to a fixed resolution of 224×224 pixels in order to meet the input specifications of deep learning models that have already been trained,
- normalization of pixel intensity values to the range **[0, 1]**.

Data augmentation strategies were performed methodically in light of the inherent diversity in hand signs caused by rotation, scale, lighting, and intra-signer differences:

- random rotations (±15°),
- horizontal flipping,
- brightness variation,
- random scaling.

The augmented dataset is formally defined as

$$D^l = \bigcup_{i=1}^{N} \bigcup_{j=1}^{M} T_j(x_i),$$

where x_i is the original image, $T_j(\cdot)$ represents augmentation transformations (rotation, flip, brightness, scaling), M is the number of augmentations per image, and N is the number of original samples. This improves the model's capacity for generalization by guaranteeing that every training period experiences various versions of the same sign (Srivastava *et al.*, 2024).

3.3. Model architecture and design

The basis of the recognition system is deep convolutional neural networks (CNNs). Six well-known models were benchmarked in order to capitalize on the complementing capabilities of different architectures:

- conventional CNN architectures,
- residual learning network, ResNet-50,
- dense connectivity network, DenseNet-121,
- deep convolutional model, VGG16,
- lightweight, mobile-optimized network, MobileNetV2,
- efficient scaling model, EfficientNet-B0.

A hybrid CNN-ResNet design was also put out, combining deep residual blocks with shallow CNN layers. Shallow layers record low-level edges and textures, whereas residual depths provide hierarchical semantic learning, allowing for thorough feature extraction in this hybrid architecture. Following concatenation and passage through completely linked layers, feature maps from both branches are classified using softmax across 26 alphabets (Damdoo, Kumar, 2025).

3.4. Training procedure and optimization

In order to minimise the difference between the real labels y and the projected probability distribution $y^{\wedge} = f(x;\theta)$, where θ stands for the trainable parameters, the models were trained using cross-entropy loss. The Adam optimiser, chosen for its adjustable learning rate characteristics that enable faster convergence and consistent gradient updates, was used to carry out the optimization. In order to achieve a balance between computational efficiency and performance, hyperparameters such as batch size, number of epochs, and learning rate were empirically changed. To lessen overfitting, early halting and dropout regularization strategies were also used.

3.5. Real-time recognition and speech synthesis pipeline

The trained hybrid CNN-ResNet model was integrated into a real-time recognition system using camera input streams to facilitate real-world usability:

- video capture: live frames are extracted from the webcam feed,
- preprocessing: each frame is resized and normalized consistently with training,
- inference: the model infers the current sign within each frame,
- output: the recognized alphabet is placed on the television display,
- TTS conversion: the recognized character is converted into natural-sounding speech via a TTS engine, providing immediate auditory feedback.

This two-way communication method effectively bridges barriers by improving accessibility for both non-signing interlocutors and hearing-impaired signers.

4. Results

The performance of the proposed ISL recognition framework was analyzed in two stages:

- static image classification using deep learning models,
- real-time recognition with text and speech integration.

The static image classification was very accurate across all ISL alphabets, indicating the efficacy of deep learning algorithms for capturing complicated hand motion patterns. Real-time recognition (Sharma, Singh, 2022) effectively translated motions into text and speech, demonstrating the system's practical usability for assisted communication. Overall, the framework showed robust performance under varying lighting and background conditions, confirming its reliability for real-world usage (Pandey *et al.*, 2025).

4.1. Static image classification results

The curated ISL dataset was used to train six baseline deep learning models (CNN, ResNet-50, DenseNet-121, VGG16, MobileNetV2, and EfficientNet-B0) and compare them to the novel CNN–ResNet architecture in order to assess the efficacy of the suggested framework. Performance metrics, which are presented in Table 2, were assessed using accuracy, precision, recall, and $F1$-score. The study demonstrates the hybrid method's superiority while outlining each model's unique benefits and drawbacks. To gain a better understanding of classification performance and model resilience, visual aids, including confusion matrices, ROC curves, and accuracy comparison charts, were used in addition to numerical measures.

Table 2. Deep learning models' comparative performance in ISL alphabet recognition.

Model	Accuracy [%]	Precision [%]	Recall [%]	$F1$-score [%]
CNN	96	90	86	86
ResNet-50	96	98	98	98
DenseNet-121	91	97	97	96
VGG16	95	99	99	99
MobileNetV2	88	90	91	90
EfficientNet-B0	91	99	99	99
CNN–ResNet (proposed)	**98**	99	99	99

The classification accuracy attained by several deep learning architectures is compared in Fig. 2. Due to their greater complexity and sensitivity to dataset size, VGG16, DenseNet-121, MobileNetV2, and EfficientNet-B0 obtained somewhat lower values than the standard CNN and ResNet-50 models, which separately achieved accuracies of about 96 %, according to the results. On the other hand, the CNN–ResNet model outperformed all baseline architectures with the maximum accuracy of 98 %. This enhancement demonstrates how well shallow convolutional filters and deep residual learning work together to capture both high-level and low-level information for reliable ISL alphabet identification.

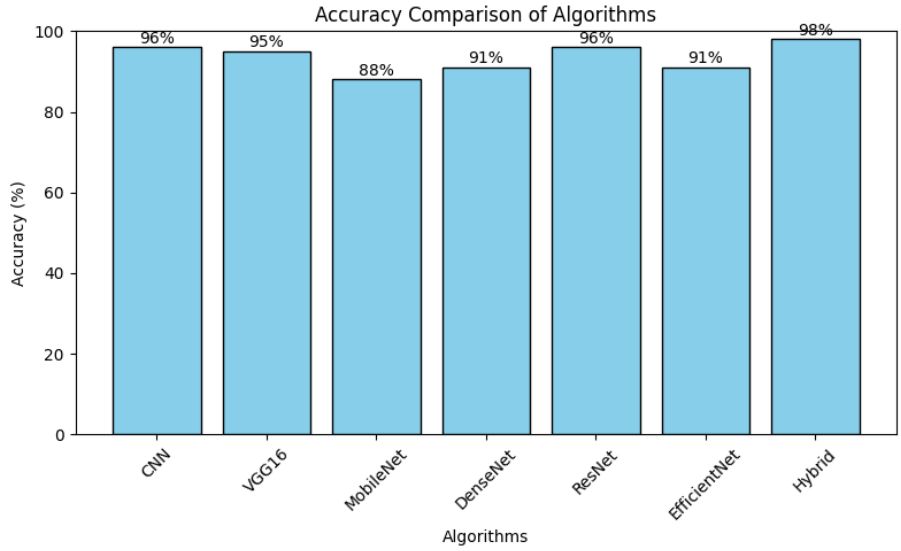

Fig. 2. Comparative analysis of the classification accuracy achieved by different deep learning architectures.

The suggested hybrid CNN–ResNet model for the ISL alphabet recognition's confusion matrix is displayed in Fig. 3. With relatively few incorrect classifications, the diagonal dominance shows that most alphabets were correctly categorized. Between visually comparable motions, such as E vs. F and P vs. R, which have overlapping hand shapes, errors were most common. However, in contrast to baseline networks, the hybrid model decreased these confusions, indicating that it can learn fine-grained differences between ISL indicators.

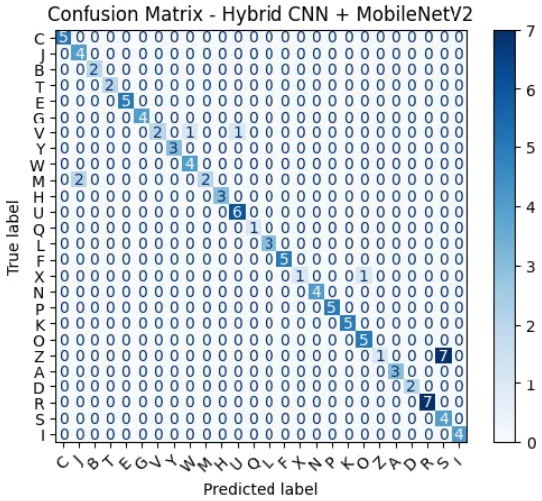

FIG. 3. Confusion matrix of the proposed hybrid CNN–ResNet model for ISL alphabet recognition.

The deep learning models' ROC curves are shown in Fig. 4 for this proposed research. The hybrid CNN-ResNet model achieved the highest area under the curve (AUC) while maintaining the optimal mix of sensitivity and specificity. In line with their lower recognition accuracy, lightweight models like MobileNetV2 and EfficientNet-B0

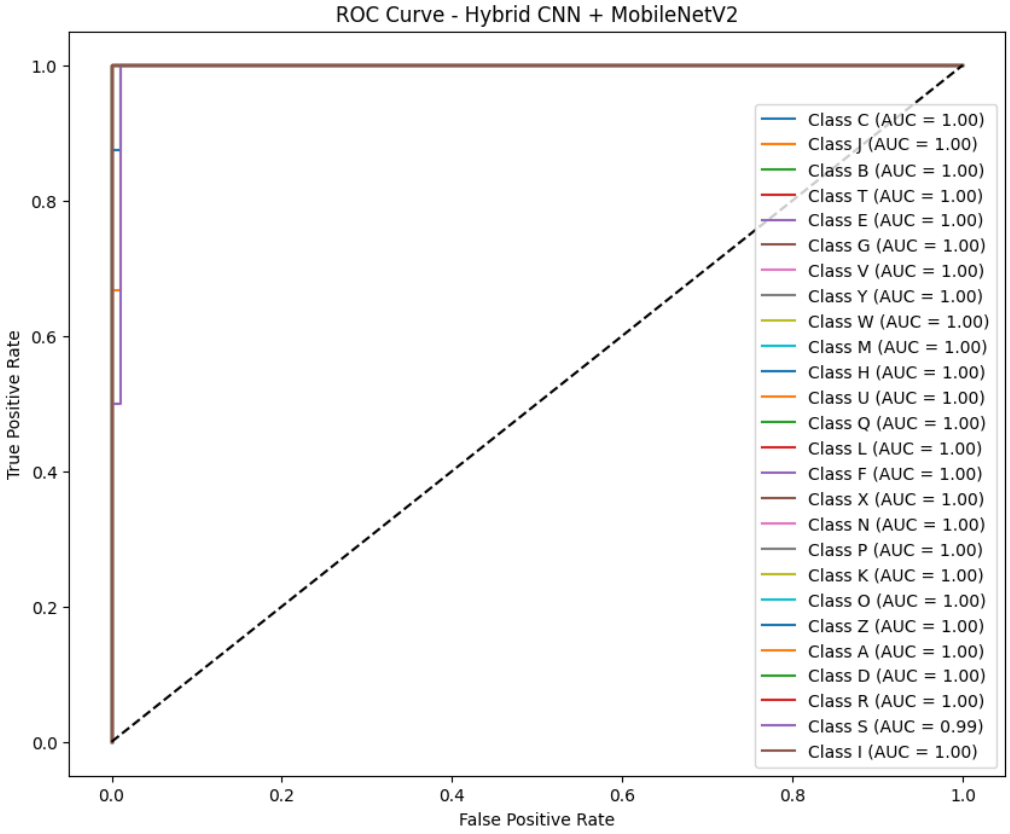

FIG. 4. Deep learning models' ROC curves.

produced somewhat lower AUC values, even if CNN and ResNet-50 also showed impressive performance. The hybrid model's generalization and robustness for ISL alphabet classification are supported by the ROC analysis.

The comparative evaluation of deep learning models for static ISL alphabet recognition showed that CNN and ResNet-50 individually achieved an accuracy of around 96 %, confirming their strong ability to extract visual features from hand gesture images. Other transfer learning models such as VGG16, DenseNet-121, MobileNetV2, and EfficientNet-B0 achieved slightly lower accuracies, largely due to the relatively small dataset size, which increased their susceptibility to overfitting. The proposed hybrid CNN–ResNet model, on the other hand, continuously beat all baselines, achieving the 98 % classification accuracy as well as better precision, recall, and $F1$ points. This improvement demonstrates the advantage of combining shallow CNN features with the deeper residual features of ResNet, resulting in more robust and discriminative feature representations for ISL alphabets.

4.1.1. Confusion matrix analysis

To gain deeper insight into the classification performance, confusion matrices were generated for each model, with a focus on the proposed CNN–ResNet architecture (see Fig. 5). The confusion matrix (Fig. 2) indicates the distribution of correctly and erroneously classified ISL alphabets.

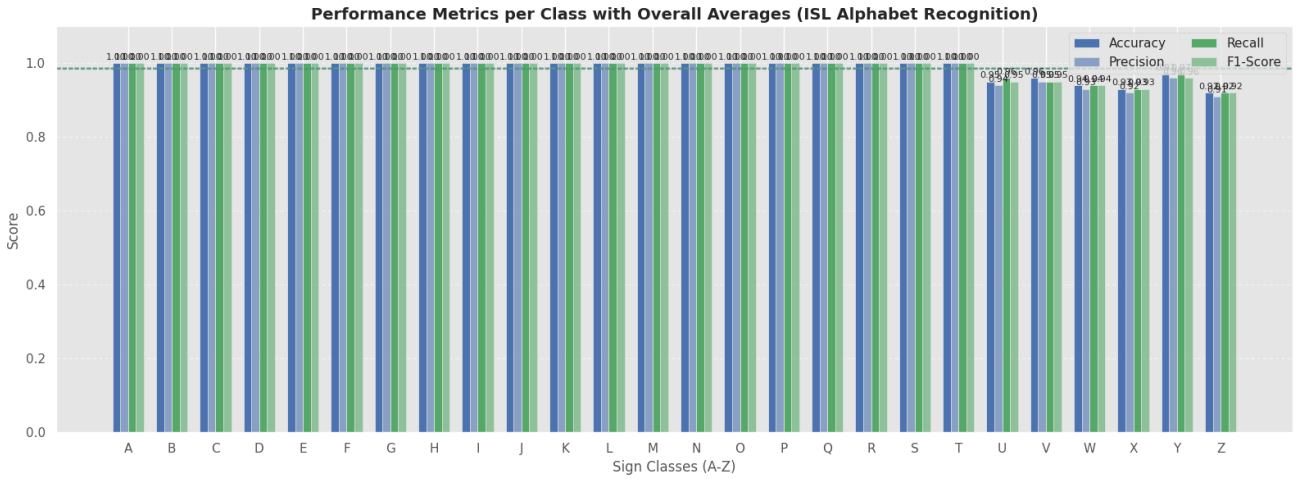

Fig. 5. Distribution of classification performance across ISL alphabet classes based on accuracy, precision, recall, and $F1$-score of the proposed CNN–ResNet model.

4.1.2. ROC curves and AUC analysis

ROC curves were plotted for each model to evaluate their trade-off between the true positive rate (TPR) and false positive rate (FPR). The AUC was computed as a quantitative indicator of model performance (Fig. 6).

Among all ISL alphabet classes, the CNN–ResNet model had the greatest AUC values, demonstrating its exceptional generalization skills. AUC values for baseline models like CNN and ResNet-50 were somewhat lower than those of the hybrid technique, but they nevertheless demonstrated outstanding performance. In Table 2, lightweight models like MobileNetV2 and EfficientNet-B0 showed comparatively lower AUC scores, which was in line with their lower recognition accuracy.

The novel CNN–ResNet architecture offers a more balanced trade-off between sensitivity and specificity, which makes it more appropriate for reliable ISL identification in practical applications, as the ROC curves verify.

4.2. Real-time recognition results

The proposed hybrid CNN–ResNet model was used for real-time ISL recognition after the static classification performance was validated. The trained model was linked to a webcam stream, which allowed for continuous TTS synthesis, preprocessing, classification, and frame recording.

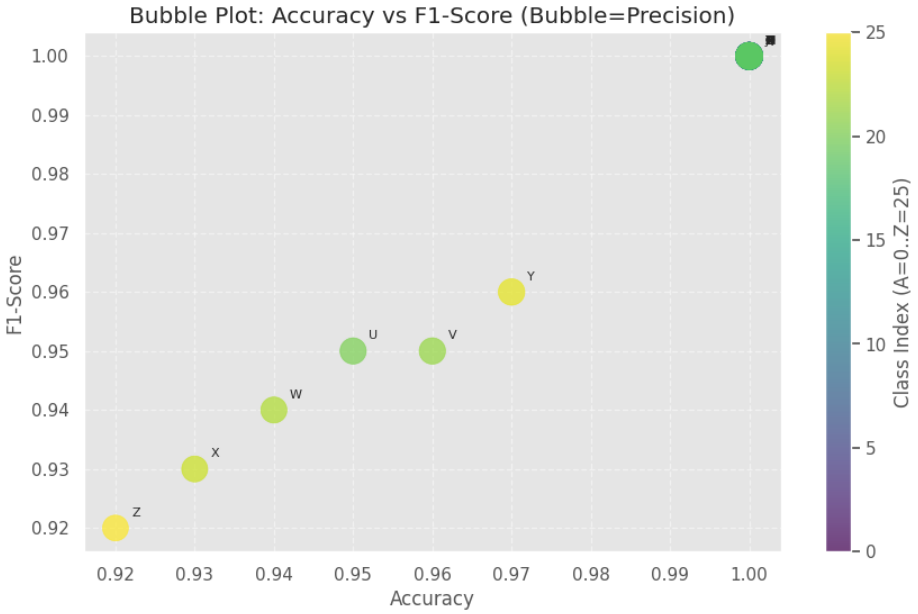

FIG. 6. Accuracy versus $F1$-score across ISL alphabet classes, with bubble size representing precision.

The system was able to identify hand motions during testing and provide the matching ISL alphabet on the screen (Fig. 7). Furthermore, real-time speech conversion of the expected output was performed (Fig. 8), providing non-signers with both visual and audible feedback.

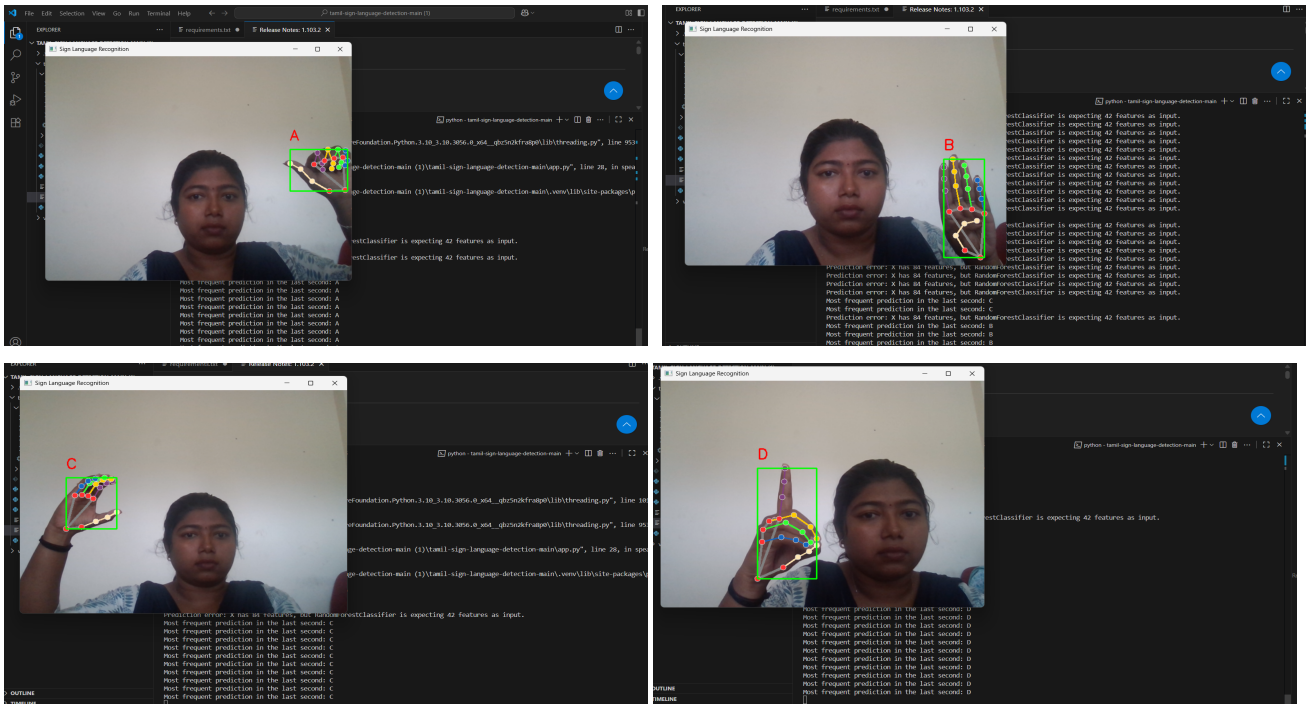

FIG. 7. Real-time recognition of ISL alphabets with text output.

The system consistently recognized alphabets and showed resilience in a range of lighting situations and signers. Although the streaming nature of the input prevented frame-wise accuracy from being calculated, qualitative testing verified that the hybrid CNN–ResNet model was highly generalizable beyond static pictures.

The per-class performance metrics for the proposed CNN–ResNet model for ISL alphabet recognition are shown in this figure: accuracy, precision, recall, and $F1$-score. Most alphabets (A–T) had near-perfect scores on

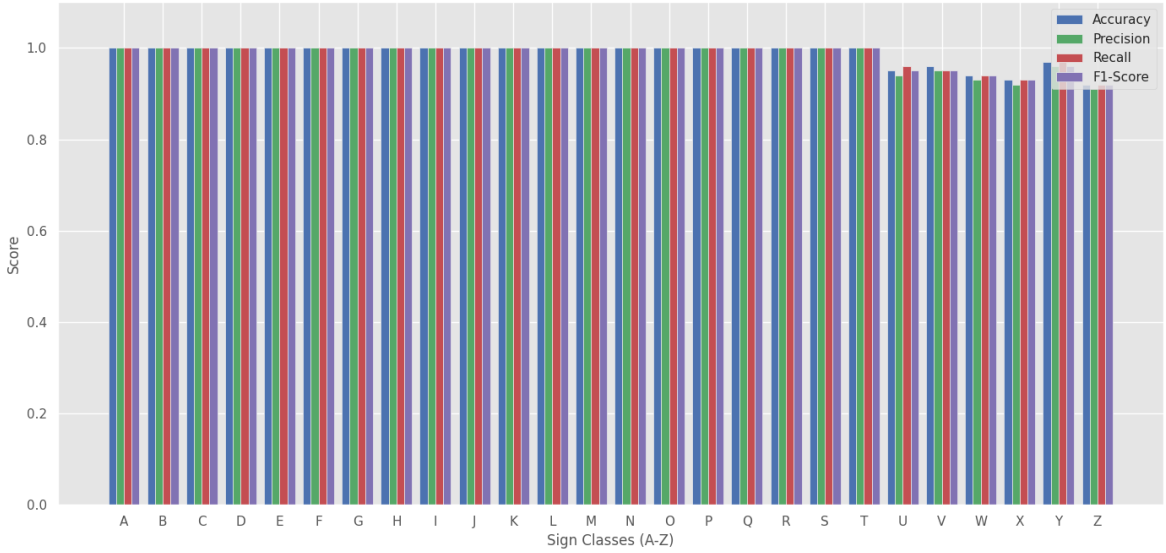

FIG. 8. Performance metrics for ISL alphabet recognition.

all measures, according to the data, demonstrating a very consistent and trustworthy categorisation. For a few visually similar alphabets, such as U, V, W, X, and Z, there is a minor performance drop.

The performance measures (accuracy, precision, recall, and $F1$-score) for each of the 26 ISL alphabets are displayed in this radar plot using a proposed hybrid CNN–ResNet model. The plot's almost perfect square shape, with values grouped near 1.0 on all axes, shows that the model consistently performs well across sign classes and assessment criteria. The radar structure shows the model's balanced and consistent performance, proving its durability and capacity to sustain high recognition accuracy without losing precision, recall, or $F1$-score. However, there are minor departures from the boundary in a few classes.

The real-time recognition performance of the hybrid CNN–ResNet model is shown in Fig. 10. The pictures are samples of ISL alphabets that were taken straight from camera input, with the signer's hand gesture and the anticipated label shown on the screen. These findings show that the model can sustain strong performance in a variety of illumination situations and signer changes while reliably identifying ISL alphabets in real-world circumstances. The system's capacity to generalize over all 26 ISL characters is demonstrated by the picture, which covers representative samples like the beginning, middle, and end of the alphabet set. Also accuracy, precision, recall, and $F1$-score are the four main metrics used to assess the performance of the suggested ISL (SANJUSARAN *et al.*, 2024) letter recognition model, as seen in the previous figures. All 26 sign classes are compared in the bar chart, where the majority of alphabets received nearly flawless scores. The performance across all classes is illustrated in Fig. 9.

4.3. DISCUSSION

The experimental assessment demonstrates that deep learning-based methods are quite successful in recognizing the alphabet in ISL (TAN *et al.*, 2024). Strong accuracies of 96 % were attained by baseline models like CNN and ResNet-50, confirming their capacity to extract discriminative features from static ISL pictures. The performance of transfer-learning architectures such as VGG16, DenseNet-121, MobileNetV2, and EfficientNet-B0 was somewhat worse, mostly because of the limitations of dataset size and their increased vulnerability to overfitting.

The hybrid CNN–ResNet model that was suggested produced the best results, with 98 % accuracy and improved $F1$-scores, precision, and recall. This demonstrates how well shallow convolutional filters, which maintain deep hierarchical features, work in conjunction with residual connections, which capture low-level edge and shape information. The confusion matrix study supports the hybrid design's ability to decrease misclassifications for visually comparable alphabets (such as E vs. F and P vs. R).

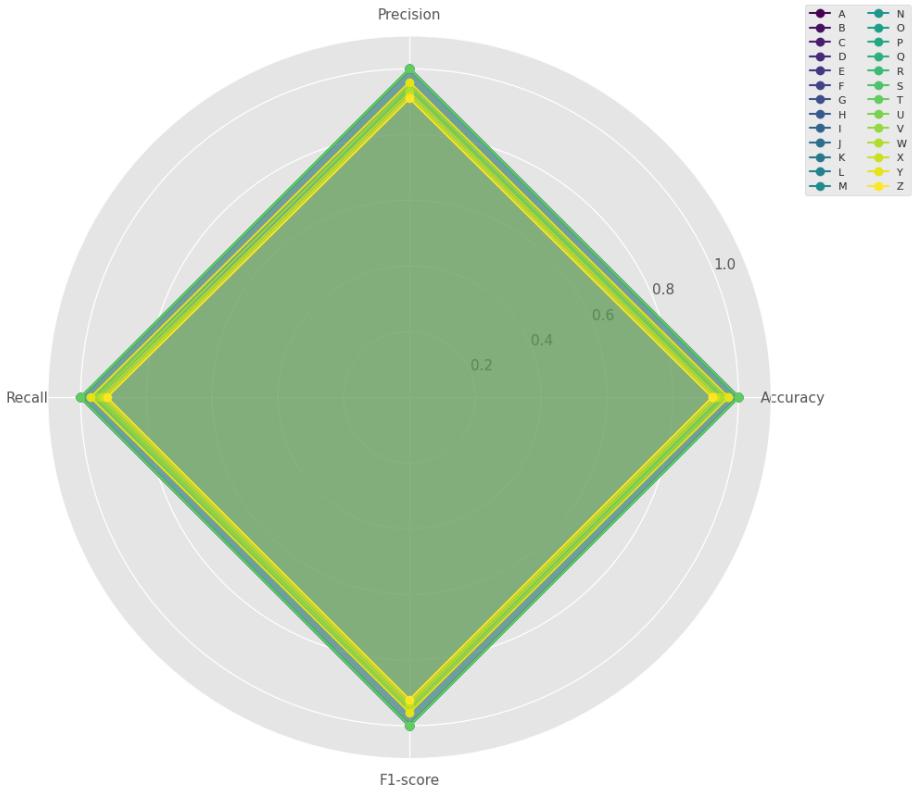

FIG. 9. Radar plot of performance metrics for all 26 ISL alphabets.

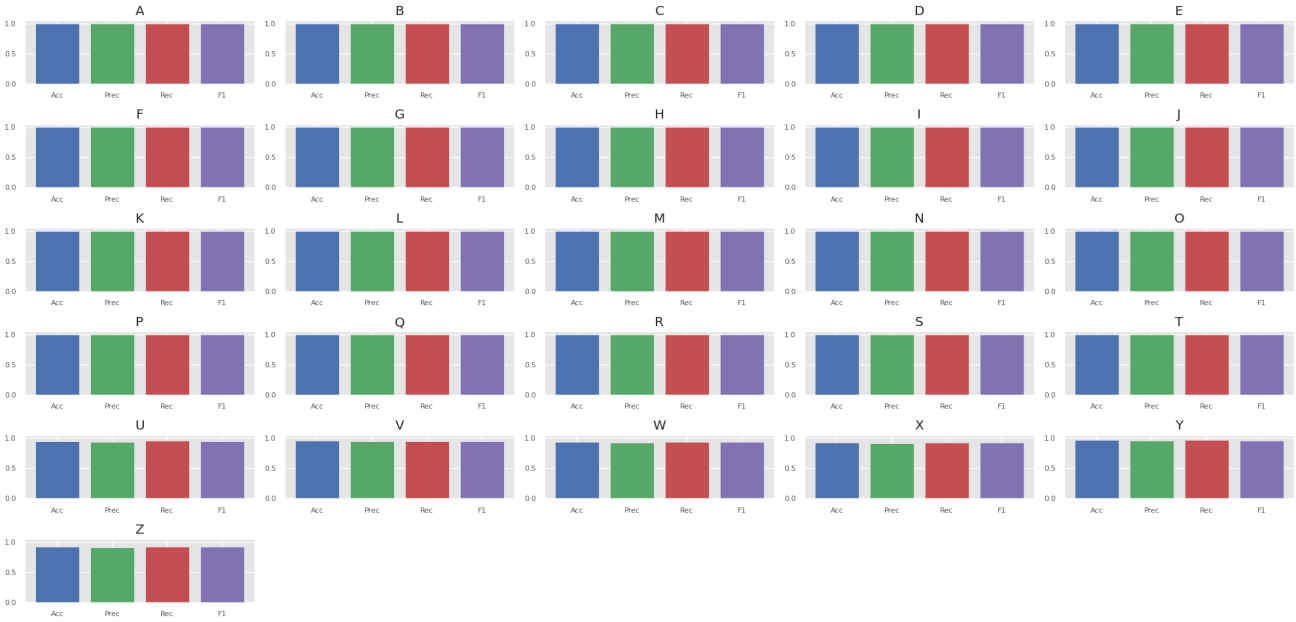

FIG. 10. Real-time recognition performance of the proposed hybrid CNN–ResNet model (small multiples: per-class metrics).

The hybrid model was shown to have the best sensitivity-specificity balance by the ROC curves, which increased its dependability for practical implementation. An important gap in ISL research was filled by integrating the system into a real-time pipeline using TTS. While the majority of earlier ISL recognition research concentrated on classifying static images, the suggested architecture allows for live interaction, allowing signed alphabets to be quickly converted into text and voice.

This dual capacity improves accessibility by giving hearing-impaired people a tool that promotes inclusive communication in education, social interactions, and human-computer interfaces.

5. Conclusion

This research addressed the dearth of real-time applications and standardized datasets in the field by presenting a deep learning-based system for ISL letter recognition. To improve variability and robustness, a curated dataset was created using the RKMVERI lexicon and the official ISL website, then preprocessed and enhanced. The hybrid CNN–ResNet model outperformed many deep learning architectures, including CNN, ResNet-50, DenseNet-121, VGG16, MobileNetV2, and EfficientNet-B0, in the evaluation.

The accuracy of the hybrid CNN–ResNet was 98 %, which was higher than the accuracy of the CNN and ResNet-50 models alone (96 %). Its enhanced discriminative capacity was validated by the confusion matrix and ROC studies, especially when dealing with visually comparable alphabets. Additionally, the system was expanded to include the TTS integration and real-time recognition pipeline, which allowed for the smooth conversion of ISL alphabets into both text and audible voice. For the community of hearing-impaired people, this invention improves accessibility and inclusion while providing useful applications in social interaction, education, and human–computer communication.

This work will be extended in the further to include continuous sign sequences, dynamic ISL gestures, and sentence-level translation. Furthermore, enhancing the model for implementation on devices with limited resources, such as smartphones or embedded systems, may allow for broad use as an inexpensive assistive technology tool.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions

Aswani Sivan conceptualized the study, curated the dataset, implemented the experiments, analyzed the results, and wrote the original draft. Chandra Eswaran provided supervision, guided the methodology design, validated the results, and critically reviewed and refined the manuscript. All authors reviewed and approved the final manuscript.

Data availability statement

The dataset created and used in this study was curated from publicly available resources (RKMVERI ISL dictionary and the ISL website). Processed data and implementation code are available from the corresponding author upon reasonable request.

Acknowledgments

The authors would like to thank the Ramakrishna Mission Vivekananda Educational and Research Institute (RKMVERI), Faculty of Disability Management and Special Education (FDMSE), Coimbatore, for access to their ISL dictionary, and the official Indian Sign Language (ISL) website for providing standardized references that supported dataset curation in this work.

References

1. Amangeldy N., Kudubayeva S., Kassymova A., Karipzhanova A., Razakhova B., Kuralov S. (2022), Sign language recognition method based on palm definition model and multiple classification, *Sensors*, **22**(17): 6621, https://doi.org/10.3390/s22176621.
2. Ashwanth B., Ventrapragada S.B., Prodduturi S.R., Depa J.R., Sharma K.V. (2023), Vision-based hand gesture recognition for Indian Sign Language using convolution neural network, *International Journal of Computer Engineering in Research Trends*, **10**(1): 1–9, https://doi.org/10.22362/ijcert/2023/v10/i01/v10i0101.
3. Chollet F. (2017), Xception: Deep learning with depthwise separable convolutions, [in:] *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, https://doi.org/10.1109/CVPR.2017.195.
4. Damdoo R., Kumar P. (2025), An integrative survey on Indian sign language recognition and translation, *IET Image Processing*, **19**(1): e700, https://doi.org/10.1049/ipr2.70000.
5. Gogoi P., Karsh B., Karsh R.K., Laskar R.H., Bhuyan M.K. (2025), Vision-based real-time gesture-to-speech translation for sign language gestures, *Procedia Computer Science*, **258**: 2050–2059, https://doi.org/10.1016/j.procs.2025.04.455.
6. Govindharajalu Kaliyaperumal V., Gopalan P.A. (2025), A deep neural network framework for dynamic two-handed Indian Sign Language recognition in hearing and speech-impaired communities, *Sensors*, **25**(12): 3652, https://doi.org/10.3390/s25123652.
7. Gupta S., Bindal A.K., Dasmana G., Shrivastva A., Sharma A. (2025), Hybrid CNN-LSTM framework for real-time deepfake detection with spatio-temporal analysis, [in:] *2025 IEEE International Conference on Smart Power, Energy, Renewables, and Transportation (SPERT)*, https://doi.org/10.1109/SPERT67079.2025.11469733.
8. He K., Zhang X., Ren S., Sun J. (2016), Deep residual learning for image recognition, [in:] *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
9. Houtsma A.J.M. (2007), Experiments on pitch perception: Implications for music and other processes, *Archives of Acoustics*, **32**(3): 475–490.
10. Huang G., Liu Z., Van Der Maaten L., Weinberger K.Q. (2017), Densely connected convolutional networks, [in:] *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, https://doi.org/10.1109/CVPR.2017.243.
11. Indian Sign Language Research and Training Center (n.d.), *Official ISL Dictionary*, https://divyangjan.depwd.gov.in/islrtc/.
12. International Organization for Standardization (1998), *Acoustics – Determination of acoustic properties in impedance tubes. Part 2: Two-microphone technique (Standard ISO No. 10534-2:1998)*, https://www.iso.org/standard/81294.html.
13. Karamanli A., Aydogdu M. (2019), Buckling of laminated composite beams due to varying in-plane loads, *Composite Structures*, **210**: 391–408, https://doi.org/10.1016/j.compstruct.2018.11.067.
14. Kingma D.P., Ba J. (2015), Adam: A method for stochastic optimization, [in:] *International Conference on Learning Representations (ICLR)*.
15. Koller O. (2020), Quantitative survey of the state of the art in sign language recognition, https://doi.org/10.48550/arXiv.2008.09918.
16. Kraśkiewicz C. *et al.* (2024), Field experiment as a tool to verify the effectiveness of prototype track structure components aimed at reducing railway noise nuisance, *Archives of Acoustics*, **49**(1): 61–71, https://doi.org/10.24425/aoa.2024.148770.
17. Mistry P., Jotaniya V., Patel P., Patel N., Hasan M. (2021), Indian sign language recognition using deep learning, [in:] *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, https://doi.org/10.1109/AIMV53313.2021.9670933.
18. Nandi U., Ghorai A., Marjit Singh M., Changdar C., Bhakta S., Pal R.K. (2022), Indian Sign Language alphabet recognition system using CNN with diffGrad optimizer and stochastic pooling, *Multimedia Tools and Applications*, **82**(7): 9627–9648, https://doi.org/10.1007/s11042-021-11595-4.
19. Pandey S., Tahseen S., Pathak R., Parveen H., Maurya M. (2025), Real-time vision-based Indian Sign Language translation using deep learning techniques, *International Journal of Innovative Research in Computer Science and Technology*, **13**(3): 38–44, https://doi.org/10.55524/ijircst.2025.13.3.6.

20. Pisharady P.K., Saerbeck M. (2015), Recent methods and databases in vision-based hand gesture recognition: A review, *Computer Vision and Image Understanding*, **141**: 152–165, https://doi.org/10.1016/j.cviu.2015.08.004.

21. Qahtan S., Alsattar H.A., Zaidan A.A., Deveci M., Pamucar D., Martinez L. (2023), A comparative study of evaluating and benchmarking sign language recognition system-based wearable sensory devices using a single fuzzy set, *Knowledge-Based Systems*, **269**: 110519, https://doi.org/10.1016/j.knosys.2023.110519.

22. Ramakrishna Mission Vivekananda Educational and Research Institute (n.d.), *Indian Sign Language Dictionary Dataset*.

23. Rastgoo R., Kiani K., Escalera S. (2021), Sign language recognition: A deep survey, *Expert Systems with Applications*, **164**: 113794, https://doi.org/10.1016/j.eswa.2020.113794.

24. Saini B., Venkatesh D., Chaudhari N., Shelake T., Gite S., Pradhan B. (2023), A comparative analysis of Indian Sign Language recognition using deep learning models, *Forum for Linguistic Studies*, **5**(1): 197–222, https://doi.org/10.18063/fls.v5i1.1617.

25. Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.C. (2018), MobileNetV2: Inverted residuals and linear bottlenecks, [in:] *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, https://doi.org/10.1109/CVPR.2018.00474.

26. Sanjusaran K., Shakthipriyan S., Supreetraju RU. (2024), A real time Indian Sign Language recognition using tensorflow, *International Journal of Engineering Research and Sustainable Technologies (IJERST)*, **2**(4): 26–33, https://doi.org/10.63458/ijerst.v2i4.98.

27. Sharma S., Singh S. (2022), Recognition of Indian Sign Language (ISL) using deep learning model, *Wireless Pers Communications*, **123**: 671–692, https://doi.org/10.1007/s11277-021-09152-1.

28. Simonyan K., Zisserman A. (2015), Very deep convolutional networks for large-scale image recognition, [in:] *International Conference on Learning Representations (ICLR)*.

29. Srivastava S., Singh S., Pooja, Prakash S. (2024), Continuous sign language recognition system using deep learning with MediaPipe Holistic, *Wireless Personal Communications*, **137**: 1455–1468, https://doi.org/10.1007/s11277-024-11356-0.

30. Tan S., Khan N., An Z., Ando Y., Kawakami R., Nakadai K. (2024), A review of deep learning-based approaches to sign language processing, *Advanced Robotics*, **38**(23): 1649–1667, https://doi.org/10.1080/01691864.2024.2442721.

31. Tan M., Le Q. (2019), EfficientNet: Rethinking model scaling for convolutional neural networks, [in:] *Proceedings of the International Conference on Machine Learning*, **97**: 6105–6114.

32. Zhou H., Zhou W., Zhou Y., Li H. (2021), Spatial-temporal multi-cue network for sign language recognition and translation, *IEEE Transactions on Multimedia*, **24**: 768–779, https://doi.org/10.1109/TMM.2021.3059098.

Archives of Acoustics Vol. 51, No. 2, pp. 359–376 (2026), https://doi.org/10.24423/archacoust.2026.4366

Technical Note

Modeling of High-speed Ultrasonic Testing of Railway Rails in Track Inspection

Sławomir Mackiewicz, Zbigniew Ranachowski*, Tomasz Katz, Tomasz Dębowski, Grzegorz Starzyński

Institute of Fundamental Technological Research, Polish Academy of Sciences
Warsaw, Poland

*Corresponding Author: zranach@ippt.pan.pl

Received October 30, 2025; revised February 17, 2026; accepted February 26, 2026; available online March 3, 2026; version of record May 5, 2026; published issue June 24, 2026.

In the paper the theoretical modeling of ultrasonic testing of railway rails with high scanning speed is considered. The model for the calculation of the ultrasonic field generated by the ultrasonic transducers and the pulse echo amplitude received after wave reflection at the defect is developed. The model is based on well-established principles of elastodynamic theory: the Rayleigh–Sommerfeld integral, the Auld reciprocity relation, and the Kirchhoff approximation. It forms the basis for design of computer program to simulate ultrasonic inspections of railway rails with automated mobile systems. The major innovation introduced in the model is taking into account the high scanning speed of the ultrasonic probes over the rail head and the limited repetition rate of the ultrasonic system. The mentioned aspects of the high-speed rail testing require the revision of one of the basic paradigms of the current ultrasonic models, which assume that the scanning speed of the ultrasonic probe is negligible in comparison to the speed of ultrasonic waves propagating in the tested material. Actually, when scanning rails at a speed of 120 km/h, the ultrasonic probe can change its position up to 5 mm between transmitting and receiving ultrasonic pulses reflected from defects located in the rail foot. Such a shift in the probe position is not negligible and should be considered in calculations. As a consequence, the ultrasonic system's slow repetition rate and fast scanning speed can make it less likely that certain rail flaws will be found. To quantitatively examine the severity of these phenomena, the new ultrasonic model and related simulation software was developed.

Keywords: non-destructive testing, railway rails defects, testing of railway rails, automatic ultrasonic testing.

1. Introduction

Due to economic pressure, there is a worldwide trend to increase train speed, traffic density, and axle loads. This has led to an increased rate of defect formation in the railway tracks and rails. The main reason for the increased number of rail defects is rolling contact fatigue (RCF) damages. The most frequent RCF defects occurring in the rail heads are head checking (HC), gauge corner cracking, squats, and transverse head cracking (tache ovale) (Kumar, 2006; International Union of Railway, 2002). In order to reduce the risk posed by rail defects, a number of non-destructive testing methods are used. Worldwide implemented rail line testing procedures include such NDT methods as visual, ultrasonic, eddy current, and magnetic (Zumpano, Meo, 2006). In the European Union, the EN 16729-1 and EN 16729-3 standards (European Committee for Standardization, 2016; 2018) are observed in this respect.

The need of permanent testing of the extensive railway network requires the implementation of fast automated inspection systems and efficient testing procedures. One of the most commonly used techniques is automated ultrasonic inspection capable of scanning the entire rail volume. An example of such an application

using the specialized inspection train was described in (Thomas *et al.*, 2007; Heckel *et al.*, 2018). The practical exploitation of the aforementioned equipment revealed that the efficient operation of the system is possible below 80 km/h. Unfortunately, this inspection speed is insufficient for some of the busiest railway lines where trains run at speeds of up to 200 km/h and more. In the dense scheduled train traffic, it is difficult to find a time window for the slow inspection trains. Therefore, the construction of inspection trains capable of performing ultrasonic inspection with speed comparable to express trains is an important and urgent matter. To investigate the possibility of high-speed ultrasonic inspection of railway rails, extensive research is necessary. The problems involved are not only of purely technical nature, i.e., connected with the construction of more robust scanners and more efficient water coupling systems, but they are of a more fundamental character. First of all, one should realize that the vast majority of ultrasonic tests performed in today's industry have a quasi-static nature. It means that the scanning speed of the tested object by ultrasonic probes is so slow that we commonly assume that during the transmitting-receiving cycle of the ultrasonic system, the probe is stationary, i.e., is in the same position during the sending and receiving of the ultrasonic pulse. But when we consider ultrasonic scanning of a railway rail with a speed of, say, 120 km/h (33 m/s), the probe position difference between sending and receiving an ultrasonic pulse reflected from the distant defect (located, for example, in the rail foot) can be about 5 mm. It is a value that cannot be neglected, as it is considerably higher than the wavelength produced by typical 2 MHz to 4 MHz ultrasonic probes used in rail inspections. In other words, the ultrasonic testing model of high-speed inspection must take into account the simultaneous propagation of the ultrasonic pulse in the examination object and the movement of the ultrasonic probe along its surface.

The other important aspect of the high-speed railway rail inspection is the inherent restriction on the system repetition rate (to c.a. 5 kHz) caused by relatively long times of flight of ultrasonic pulses reflected from defects located in the rail neck or foot. Such a restriction of the repetition rate, relatively unimportant for standard ultrasonic testing, causes substantial problems for high-speed inspections of railway rails. Specifically, for ultrasonic testing carried out with a speed of 120 km/h and a system repetition rate of 5 kHz, the scanning step (the distance between successive emissions of ultrasonic pulses along the rail length) is equal to 6.7 mm. If a small defect is located between successive emission points, it can be missed or detected with considerably smaller echo amplitude than would be detected at a standard (slow) scanning speed. Therefore, a new model of the ultrasonic inspection is needed to take into account the specific problems of ultrasonic testing with high scanning speed.

The paper presents development of an ultrasonic pulse-echo testing model which takes into account the dynamic nature of high speed ultrasonic inspection. It can be considered as an extension of our previous quasi static model presented in (Katz *et al.*, 2021) which was adequate for modelling of ultrasonic tests of railway rails conducted at low scanning speeds.

The model forms the theoretical basis for a computer program capable of simulation of high speed ultrasonic inspection of railway rails performed with different types of ultrasonic probes, including shear wave angle beams probes and longitudinal wave normal beam ones. The program allows for calculation of ultrasonic field generated by the probe depending on its parameters (frequency, refraction angle, transducer shape and size, wedge material, etc.) as well as calculation of the amplitude of pulse echo reflected from simple model defects (circles or rectangles) of arbitrary size, location, and orientation.

In the last part of the paper, some calculation results for typical rail testing configurations are presented and discussed in the context of high scanning speed. Finally, some conclusions, which may be important for designers of high-speed ultrasonic testing systems, are presented.

2. Dynamic model of ultrasonic testing with high scanning speed

As already mentioned, the commonly used model of ultrasonic testing is quasi-static, i.e., it assumes that during the transmitting-receiving cycle, the ultrasonic probe remains in the same position. Thus, any head movement associated with object scanning does not affect the amplitude of the recorded ultrasonic echoes. In such a theoretical approach, the test is carried out in a kind of virtual jump from one wave emission point to another, with the ultrasonic probe effectively frozen at each of these points during the transmission-reception cycle. In case

of high scanning speed, it is necessary to move away from this simplified picture and introduce a dynamic model of ultrasonic inspection that takes into account the simultaneous movement of the ultrasonic pulse and the scanning probe. The scheme of small defect detection in such a model is shown in Fig. 1.

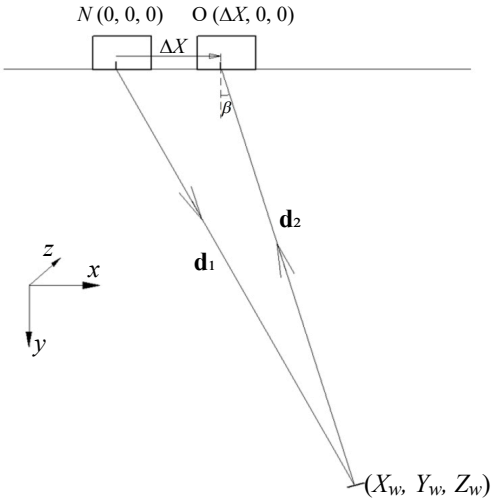

FIG. 1. Defect detection in the dynamic model of ultrasonic testing.

The transmission-reception cycle of the ultrasonic system in the pulse-echo mode can be described as follows. The probe sends an ultrasonic pulse into the tested object in the transmitting position N, which can be described by initial coordinates (0, 0, 0) in the stationary coordinate system (X, Y, Z) associated with the rail. The X-axis is parallel to the rail axis and is also a scanning axis of the ultrasonic probe; the Y-axis is directed down the rail, and the Z-axis is in the transverse direction. The emitted pulse propagates in the rail material at the speed of ultrasonic wave $\mathbf{V}_u$ and eventually arrives at a material defect located at the point with coordinates (X_w, Y_w, Z_w). The distance d_1 that the ultrasonic pulse travels during this time can be calculated from the geometry of the problem shown in the drawing, remembering, however, that the problem is essentially 3D, i.e., the defect shown does not have to be located in the plane of the drawing:

$$d_1 = \sqrt{X_w^2 + Y_w^2 + Z_w^2}. \tag{1}$$

Consequently, the time t_1 in which the ultrasonic pulse travels from the point of introduction to the defect center can be calculated using:

$$t_1 = d_1/V_u. \tag{2}$$

While the ultrasonic pulse is propagating to the defect and back, the probe moves along the rail at the scanning speed V_s, covering the distance $\Delta X = V_s(t_1 + t_2)$ in that time. However, the distance ΔX cannot be easily calculated because we do not know the time t_2 in which the pulse travels the unknown return distance d_2 from the defect to the probe. Regarding Fig. 1, one can see that the two sides of the triangle shown in the figure, ΔX and d_2, are dependent on the unknown time t_2 as follows:

$$\Delta X = V_s(t_1 + t_2), \qquad d_2 = V_u t_2. \tag{3}$$

However, all sides of the triangle shown are related by the law of cosines, giving an equation from which the unknown value of t_2 can be calculated:

$$d_1^2 = \Delta X^2 + d_2^2 - 2\Delta X d_2 \cos(90^\circ + \beta). \tag{4}$$

It should be noted that the angle $(90^\circ + \beta)$ lies in the plane of the triangle, not in the plane of the drawing. Using known trigonometric identities, Eq. (4) can be transformed into the form:

$$d_1^2 = \Delta X^2 + d_2^2 + 2\Delta X d_2 \sin\beta, \tag{5}$$

where the unknown value of $\sin\beta$ can be expressed using:

$$\sin\beta = (X_w - \Delta X)/d_2. \tag{6}$$

Substituting this value into Eq. (5), we get:

$$d_1^2 = \Delta X^2 + d_2^2 + 2\Delta X X_w - 2\Delta X^2. \tag{7}$$

Then, by substituting the values of ΔX and d_2 expressed by Eq. (3), we finally obtain the quadratic equation for the unknown transition time t_2 from the defect to the probe at its receiving position:

$$\left[V_u^2 - V_s^2\right] t_2^2 + \left[2V_s X_w - \frac{2V_s^2 d_1}{V_u}\right] t_2 + \left[\frac{2V_s d_1 X_w}{V_u} - d_1^2 - \frac{V_s^2 d_1^2}{V_u^2}\right] = 0. \tag{8}$$

By taking the positive root of the above quadratic equation and leaving only the linear terms in the V_s/V_u ratio, we obtain the approximate time of passage t_2 from the defect center to the probe in its receiving position:

$$t_2 = \frac{d_1}{V_u} - 2\frac{V_s}{V_u}\frac{X_w}{V_u}. \tag{9}$$

Knowing times t_1 and t_2, one can easily calculate the probe displacement ΔX during the time of flight of the ultrasonic pulse to and from the considered defect:

$$\Delta X = V_s(t_1 + t_2) = 2V_s\left(\frac{d_1}{V_u} - \frac{V_s}{V_u}\frac{X_w}{V_u}\right), \tag{10}$$

where d_1 is calculated from Eq. (1). It should be noted that the displacement ΔX between the points of sending and receiving the ultrasonic pulse depends on both the scanning speed V_s and the location of the defect in the rail, in the defined coordination system. The further away the defect is from the probe, the greater the distance between the points of sending and receiving of the reflected ultrasonic pulse.

The presented dynamic model of ultrasonic inspection is not fully strict as it assumes that the ultrasonic probe is effectively frozen during the acts of emitting and receiving ultrasonic pulses. Still, the time of introduction/reception of ultrasonic pulses to the material is very short (c.a. 2 μs for standard 2 MHz ultrasonic probes) as compared to the typical times of flight of ultrasonic pulses reflected from the rail defects (from 20 μs to 150 μs). It means that the model takes into account the major part of the phenomena related to the dynamic nature of the high-speed ultrasonic testing of railway rails. One can also object that the above consideration is not valid for large defects whose size is comparable to the distance between the testing probe and the defect center. Actually, it is not a big problem, as we can divide a large defect into many small fragments and calculate ΔX for each fragment separately. Then the echo amplitude is calculated as the sum of the contributions from all fragments of the larger defect.

The above solution is the basis for the modification of the calculation algorithms implemented in the SymUT software, prepared by the first author of the paper. The program calculates ΔX values using Eq. (10) based on the entered data on the scanning speed V_s and model defect position relative to the testing probe. Simulations of ultrasonic echo envelopes during rail scanning, performed applying the SymUT software is presented in Sec. 7.

3. Basic assumptions and simplifications adopted in the ultrasonic model

The model assumptions were thoroughly adjusted to the actual conditions of ultrasonic testing of railway rails. Specifically, the tested material was assumed to be a homogeneous, isotropic, elastic solid characterized by ultrasonic velocities, mass density, and attenuation coefficients typical of railway steel. The attenuation coefficient

values for longitudinal and transversal waves in rail steel were inferred from the tables given in (Ono, 2020a; 2020b). The probe wedge material was assumed to be a fluid medium with a density and longitudinal wave velocity corresponding to actually used solid materials (PMMA or Rexolite). This simplification allowed for the avoidance of unnecessary mathematical complexity in situations where only longitudinal wave propagation in probe wedges is considered.

It was assumed that the boundary between the probe wedge and the tested material is perfectly flat, neglecting the slight curvature of the running surface of the rail heads. Additionally, it was assumed perfect acoustic coupling between the probe wedge and the tested material (smooth contact boundary conditions sometimes also called slip boundary conditions). They assume continuity of normal stress and displacement and vanishing of tangential stresses at the border. The potential influence of the coupling layer thickness on the echo amplitude could be studied separately using the other model described in (Mackiewicz *et al.*, 2024).

Taking into account the considerable size of the tested items (rails), it is assumed that the modeled defects are generally situated in the far fields of ultrasonic probes, i.e., in the region where the ultrasonic field can be locally approximated by the plane wave. On the other hand, the paraxial approximation was avoided as many rail defects are cracks with unfavorable orientation (Deschamps, 1972), which may be better detected by the side rays significantly deviated from the beam axis. In the proposed model, we also assume that the ultrasonic pulses generated by the probe transducers are relatively short, as is actually the case in modern commercial ultrasonic probes. This means that the −6 dB bandwidth of simulated probes should not be smaller than 30 %. The first part of the model is a method for the calculation of the ultrasonic field generated in the tested object by the ultrasonic transducer. Following this basic functionality, the method for calculation of pulse echo amplitude from simple model defects (circle, square, or rectangular cracks) was developed based on the Auld reciprocity principle and the Kirchhoff approximation (Darmon, Chatillon, 2013).

4. Calculation of ultrasonic field generated by the ultrasonic transducer

The method of calculation of the ultrasonic field generated by the piezoelectric transducer in the tested material is essentially the same as in the standard quasi-static model. This is because we assumed that the act of introduction of ultrasonic pulse from the probe to the material is short enough that the probe can be considered static during this period of time. After emission, the pulse 'forgets' about the sending probe, and it does not matter if the probe is static or travels along the scanning surface with a high speed. So the ultrasonic field generated in the material is independent of this feature and can be calculated in the same way as for stationary probes.

Further, the standard geometrical configuration used for ultrasonic testing of railway rails with angle beam shear wave probes is considered. The calculation for longitudinal wave probes is completely analogue. Because we assumed that the act of introducing the ultrasonic pulse to the material is very short (compared to the total time of flight in the tested rail) the calculation of the ultrasonic field generated in the material by the moving ultrasonic probe may be performed in the same way as for stationary probe. The probe is coupled to the flat surface of the tested object with a thin layer of coupling medium. The piezoelectric transducer is attached to the refracting wedge made of a formally fluid medium with characteristics (density and longitudinal wave velocity) compatible with plastic material actually used for the fabrication of the refracting wedges (PMMA, polystyrene, or Rexolite). This way the wedge material's shear stiffness can be neglected. This assumption is also compatible with the fact that the probe wedge is coupled to the tested material with a thin layer of liquid medium which does not transmit shear stresses.

In line with a common practice in ultrasonic modeling, we assume piston-like transducer vibrations with an angular frequency ω and a uniform particle velocity v_0 over its front surface. The ultrasonic wave propagating in the wedge hits the boundary between the wedge and the tested material at an angle which is the same as the wedge angle α. The wedge angle is selected between the 1st and 2nd critical angles so that only one refracted wave (T-type) is generated in the tested material.

The refraction angle of the transversal wave β is related to the incidence angle of the longitudinal wave α through Snell's law:

$$\frac{\sin\alpha}{\sin\beta} = \frac{V_{L1}}{V_{T2}}, \tag{11}$$

where V_{L1} is the velocity of L-type wave in the wedge material and V_{T2} is the velocity of T-type wave in the tested material. Calmon *et al.* (1998) showed that it is possible to use the Rayleigh–Sommerfeld integral for the calculation of the ultrasonic field in the wedge material. It expresses the acoustic pressure p as an integral over the radiating transducer surface:

$$p(\mathbf{x}) = \frac{-i\omega v_0 \rho_1}{2\pi} \int\limits_{S_t} \frac{e^{ikr}}{r} \mathrm{dS}, \tag{12}$$

where ρ_1 is the mass density of wedge material, ω is the angular frequency of ultrasonic vibration, S_t is the surface of the transmitting transducer, v_0 is the normal particle velocity at the transducer face, k is the wavenumber of ultrasonic wave in the wedge, and r is the distance between the field point, and $\mathbf{x}$ is the current integration point on the transducer surface. Using the integral Eq. (12) one could easily calculate the ultrasonic field in the wedge material, but not in the tested material. In order to accomplish this more difficult task, we used the so-called pencil model introduced by Deschamps (1972), for calculations of electromagnetic beams generated by radars. This concept was also used for the modelling of ultrasonic waves by Raillon, Lecoeur-Taïbi (2000) and Gengembre, Lhemery (2000). The discussed model has been proved to be very effective in terms of calculation accuracy and computing efficiency. Finally, the ultrasonic field at the point $\mathbf{x}$ in the tested material is described by the particle velocity amplitude $\mathbf{v}(\mathbf{x}, \boldsymbol{\omega})$ given in the form:

$$\mathbf{v}(\mathbf{x}, \boldsymbol{\omega}) = \frac{-i\omega v_0}{2\pi V_{L1}} \int\limits_{Sr} \frac{T_{12}^v(\alpha) e^{i(k_1 r_1 + k_2 r_2)}}{\left(r_1 + \frac{V_{T2}}{V_{L1}} \frac{\cos^2\alpha}{\cos^2\beta} r_2\right)^{1/2} \left(r_1 + \frac{V_{T2}}{V_{L1}} r_2\right)^{1/2}} \mathrm{dS}, \tag{13}$$

where V_{L1} is the velocity of the L-type wave in the wedge material, V_{T2} is the velocity of the T-type wave in the tested material, α is the incidence angle of a pencil central ray at the wedge-material boundary, β is the refraction angle of a pencil central ray at the wedge-material boundary, r_1 is the section of the pencil central ray in the wedge material, r_2 is the section of the pencil central ray in the tested material, $k_1 = \omega V_{L1}$ is the wave number of the L-type wave in the wedge material, $k_2 = \omega / V_{T2}$ is the wave number of the T-type wave in the tested material, S_r is the surface of the radiating transducer, and T_{12}^v is the particle velocity transmission coefficient at the wedge-material boundary.

Equation (13) sums up the contributions from partial waves emanating from all vibrating points of the transducer face. The integral sums up the scalar values of particle velocity amplitudes, neglecting the fact that partial waves coming from different points of the transducer surface have particle velocities with slightly different directions. This is a typical simplification in the scalar theories of diffraction, and its actual meaning in the considered case is illustrated in Fig. 2.

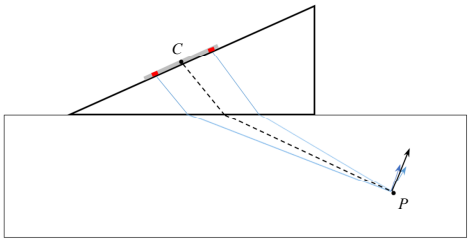

Fig. 2. Polarizations of partial waves coming from different points of vibrating transducer, placed at the top of the wedge.

It can be seen that in the far field, the differences between polarizations of partial waves coming from different parts of vibrating transducers are small, and the simplification made in Eq. (13) causes only a slight

overestimation of the calculated amplitude of the particle velocity of the generated ultrasonic wave. Such scalar values may be conveniently displayed on graphs illustrating the distribution of ultrasonic fields generated by different probes. For more advanced simulations, and specifically for the calculation of ultrasonic echo amplitudes from model defects, we need to know not only a wave amplitude but also its polarization. Based on heuristic reasoning, without a rigorous proof, we assume that the polarization of the ultrasonic wave generated by the whole vibrating transducer at a given point P is the same as the polarization of the partial wave emanating from the transducer center and coming to the same point P (see Fig. 2). This assumption, together with Eq. (13), gives the full description of the ultrasonic field generated in an isotropic solid by an angle beam shear wave probe.

All factors included in Eq. (13) can be readily calculated if we can determine α, β, r_1, and r_2 for any combination of the source point on the transducer surface and the evaluation point in the tested material. This problem is equivalent to determining the Fermat path between two given points in two bordering materials. It can be easily solved using a numerical algorithm looking for the minimum time of flight of ultrasonic waves between two given points.

The integral Eq. (13), gives the distribution of the ultrasonic field of a monochromatic wave with a circular frequency ω. To calculate the amplitude of short pulses generated by the real ultrasonic probes, we need to calculate $v(\mathbf{x},\omega_k)$ for several discrete frequencies ω_k contained in the frequency spectrum of the transmitting transducer, then multiply them by the corresponding amplitudes of the spectrum, and finally perform the inverse fast Fourier transform. This way we can calculate the pulse shape in the time domain and finally obtain its maximum amplitude.

5. Calculation of pulse echo amplitude from model defects

Calculation of the ultrasonic field generated by the ultrasonic probe in the tested material is a very important aspect of the theoretical modeling of ultrasonic NDT inspection. It allows checking if potential defects are within the sensitivity zone of the selected probes and precisely adjusting its basic parameters (central frequency, transducer size, and refraction angle) to properly cover the inspection zone. However, illuminating the defect with a strong ultrasonic beam is not a sufficient condition for its detection in ultrasonic examination. The shape, orientation, and size of the defect are also of great importance. In the case of railway rails, most defects are relatively large cracks with varied size and orientation. To reasonably simulate such types of defects, we introduced model defects in the form of circles, squares, and rectangles of defined size and orientation. We also assume that the model defect surface is flat and stress-free. To calculate the amplitude of ultrasonic echo reflected from the model defect, we make use of the reciprocity theorem presented by Auld (1979) and modified by Schmerr (2016), which is more suitable for NDT applications. The reciprocity theorem greatly reduces the complexity of the calculation of ultrasonic pulse echo amplitude in comparison to the direct approach requiring successive calculations of the ultrasonic field generated by the transmitting transducer incident on the defect, then assessment of the scattered field from the defect, and finally the integration of this field over the surface of the receiving transducer.

The general formula derived by Schmerr (2016) gives the voltage amplitude $V_R(\omega)$ generated on the receiving transducer by the ultrasonic wave reflected from the model defect. It is a definite integral over the surface of the model defect of the products of stress and particle velocity components of two elementary ultrasonic solutions:

$$V_R(\omega) = \frac{\mathrm{SF}(\omega)}{\rho_1 V_{L1} S_T v_T^{(1)} v_R^{(2)}} \int\limits_{S_d} \left(\tau_{ij}^{(1)} v_j^{(2)} - \tau_{ij}^{(2)} v_j^{(1)}\right) n_i \mathrm{dS}, \tag{14}$$

where $V_R(\omega)$ is the voltage generated on the receiving transducer by ultrasonic waves reflected from the model defect, $v_j^{(1)}$ is the component of particle velocity for solution (1), $v_j^{(2)}$ is the component of particle velocity for solution (2), $\tau_{ij}^{(1)}$ is the component of stress tensor for solution (1), $\tau_{ij}^{(2)}$ is the component of stress tensor for solution (2), n_j is the component of the unit vector $\mathbf{n}$ normal to the model defect, $v_T^{(1)}$ is the amplitude of

transmitting transducer vibrations for solution (1), $v_R^{(2)}$ is the amplitude of receiving transducer vibrations for solution (2), S_T is the surface area of the transmitting transducer, S_d is the surface of the model defect, $\mathrm{SF}(\omega)$ is the system function, which incorporates the total effect of all electrical and electromechanical components of the ultrasonic system.

Solution (1) refers to the ultrasonic field generated in the tested material by the transmitting transducer, which physically interacts with the model defect and reflects at its surface. Actually, it is the field we considered in Sec. 4, which can be calculated using Eq. (13). Solution (2) refers to an imaginary (nonexistent) ultrasonic field that would be generated in the tested material if the receiving transducer acted as a transmitting transducer. The ultrasonic field calculated in the solution (2) neglects the presence of the defect and can be conveniently interpreted as the 'sensitivity field' of the receiving transducer. In our dynamic model of ultrasonic testing, the solution (2) is generated by the same transducer but shifted on the x-axis by the ΔX value calculated using the Eq. (3) and Eq. (8).

The general form of the reciprocity Eq. (14) cannot be directly used in practical modeling of ultrasonic inspection because it is dependent on some factors that are impossible or very difficult to determine based on information available to NDT personnel. This applies in particular to the $\mathrm{SF}(\omega)$ and vibration amplitudes of transmitting and receiving transducer – $v_T^{(1)}$ and $v_R^{(2)}$. We can simplify the Eq. (14) considering what is really necessary for modeling a real ultrasonic examination and what is unnecessary or excess information.

First of all, in NDT applications we do not need the absolute values of voltage generated on the receiving transducer $V_R(\omega)$ but only relations between voltage amplitudes received from modeled defects and from the defined reference reflector. This relation is commonly expressed on a logarithmic scale in decibels [dB]. For this reason, in Eq. (14), we can ignore all constant values appearing before the integral sign. The $\mathrm{SF}(\omega)$ cannot be neglected because this function carries important information about the bandwidth of the ultrasonic system. Principally, this function should be determined specifically for a given ultrasonic system (consisting of transmitter, receiver, probes, and cabling) using a special calibration procedure described by Schmerr (2016). Unfortunately, this procedure is rather difficult to perform, especially at the project design phase when the individual components of the modeled ultrasonic system may not be physically available. To overcome this difficulty, we propose a simplified approach based only on information readily accessible from standard specifications of ultrasonic equipment.

Considering that probes used in ultrasonic testing of railway rails have frequencies within the limited range of 1 MHz to 5 MHz, we can safely assume that the typical ultrasonic receiver, in this range, has an almost flat frequency characteristic. The same can be said about the frequency band of a typical ultrasonic transmitter operating in a standard spike mode. Then we can safely assume that the shape of the $\mathrm{SF}(\omega)$ is mainly determined by the frequency spectrum of the ultrasonic probe. According to International Organization for Standardization (2020), the frequency spectrum of an ultrasonic probe is defined by the −6 dB frequency band measured in specified standard conditions. For commercial probes the parameter called relative frequency bandwidth (BW) is usually provided in the technical data sheets and is easily available to NDT personnel.

Based on the probe BW parameter and the probe central frequency (f_0) we can approximately determine the SF in the form of the Gaussian function:

$$\mathrm{SF}(\omega) = e^{-\frac{(\omega/\omega_0)^2}{2\sigma^2}}, \tag{15}$$

where $\omega_0 = 2\pi f_0$ and $\sigma = \mathrm{BW}/235$. The shape of the Gaussian function reasonably resembles the typical shape of ultrasonic transducer frequency characteristics, and the σ parameter was chosen so that the half-width of this function is equal to the −6 dB bandwidth of the modeled transducer. Although Eq. (15) is only a rough approximation of the actual SF, it takes into account the most important factor determining the ultrasonic system frequency characteristic, that is the bandwidth of the ultrasonic probe.

Another simplification that can be applied to Eq. (14) follows from the fact that we are considering crack-like model defects. It means that the defect surfaces are stress-free, and in Eq. (14), we can put $\tau_{ij}^{(1)} \equiv 0$. After all the aforementioned simplifications, we can rewrite Eq. (14) in a much simpler form:

$$V_R(\omega) = -\mathrm{SF}(\omega) \int_{S_d} \tau_{ij}^{(2)} v_j^{(1)} n_i \mathrm{d}S, \tag{16}$$

where the SF(ω) is given by Eq. (15) and the minus sign is irrelevant because in our model all constants before the integral sign can be ignored.

Now the calculation of components of the stress tensor $\tau_{ij}^{(2)}$ on the surface of the model defect is presented. These stress components are not zero because solution (2) in the reciprocity formula ignores the existence of defects in the tested material. So, we just need to calculate the ultrasonic field of the receiving transducer for the defect-free elastic solid and express it in the form of stress tensor components.

The general stress-strain relation for elastic solid is:

$$\tau_{ij}^{(2)} = C_{ijkl} e_{kl}^{(2)} = C_{ijkl} \frac{1}{2} \left(\frac{\partial u_k^{(2)}}{\partial x_l} + \frac{\partial u_l^{(2)}}{\partial k} \right) = C_{ijkl} \frac{\partial u_k^{(2)}}{\partial x_l}, \tag{17}$$

where $u_k^{(2)}$ is the component of the displacement vector for solution (2), $e_{kl}^{(2)}$ is the component of the strain tensor for solution (2), C_{ijkl} is the component of the stiffness tensor for tested material.

For isotropic solid, components of the stiffness tensor can be expressed in much simpler form:

$$C_{ijkl} = \lambda \delta_{ij} \delta_{kl} + \mu \left(\delta_{ik} \delta_{jl} + \delta_{il} \delta_{jk} \right), \tag{18}$$

where λ, μ are the Lame elastic constants for the tested material, δ_{kl} is the delta Kronecker symbol.

Substituting Eq. (18) to Eq. (17) we obtain:

$$\tau_{ij}^{(2)} = \lambda \delta_{ij} \frac{\partial u_k^{(2)}}{\partial x_k} + \mu \left(\frac{\partial u_i^{(2)}}{\partial x_j} + \frac{\partial u_j^{(2)}}{\partial x_i} \right). \tag{19}$$

In the next step, the expressions for partial derivatives of the components of the displacement vector appearing in Eq. (19), would be determined. At this step the initial assumption that the ultrasonic field generated by the probe at the model defect may be locally approximated by the plane wave can be implemented. In the considered case, it is the plane wave of transversal type with direction vector $\mathbf{e}^{(2)}$, polarization vector $\mathbf{d}^{(2)}$, and wave number k_2. The geometrical configuration for solutions (1) and (2) is shown in Fig. 3.

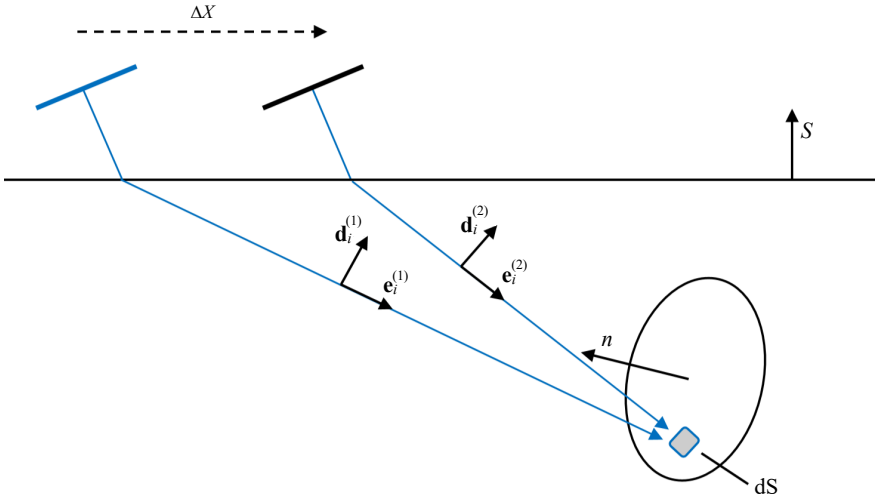

FIG. 3. Illustration supporting calculation of pulse-echo amplitude using reciprocity principle.

In this case, it can be put:

$$\mathbf{u}^{(2)} = \frac{1}{-i\omega} \mathbf{d}^{(2)} v^{(2)} (\mathbf{x}, \omega), \tag{20}$$

where $\mathbf{u}^{(2)}$ is the displacement vector of the local plane wave of solution(2), $\mathbf{d}^{(2)}$ is the polarization unit vector of the local plane wave of solution(2), $v^{(2)}(\mathbf{x}, \omega)$ is the amplitude of the local plane wave approximating solution (2).

Consequently, the partial derivatives of the displacement field (Eq. (20)) can be expressed as

$$\frac{\partial u_i^{(2)}}{\partial x_j} = \frac{ik_2}{-i\omega} d_i^{(2)} e_j^{(2)} v^{(2)}(\mathbf{x},\omega) = \frac{-1}{V_T} d_i^{(2)} e_j^{(2)} v^{(2)}(x,\omega), \tag{21}$$

where k_2 is the wavenumber for the transversal wave in tested material and V_T is the velocity of the transversal wave in tested material.

Substituting Eq. (21) into Eq. (19) we obtain:

$$\tau_{ij}^{(2)} = v^{(2)}(\mathbf{x},\omega) \frac{-1}{V_T} \left[\lambda \delta_{ij} d_i^{(2)} e_j^{(2)} + \mu \left(d_i^{(2)} e_j^{(2)} + d_j^{(2)} e_i^{(2)}\right)\right]. \tag{22}$$

The first component in square brackets is a scalar product of the polarization and directional vectors of the shear wave and must identically equal zero. Consequently, the final expression for stress components of solution (2) discussed in Eq. (14), Eq. (16), Eq. (17), and Eq. (19) takes the form:

$$\tau_{ij}^{(2)} = \frac{-\mu}{V_T} v^{(2)}(\mathbf{x},\omega) \left(d_i^{(2)} e_j^{(2)} + d_j^{(2)} e_i^{(2)}\right). \tag{23}$$

The scalar amplitude $v^{(2)}(\mathbf{x},\omega)$ of the particle velocity can be calculated from the integral Eq. (13) by replacing the transmitting transducer with the receiving transducer shifted from the initial position $(0,0,0)$ to the receiving position $(\Delta X,0,0)$, where ΔX must be calculated from Eq. (10). The directional vector $\mathbf{e}^{(2)}$ can be calculated based on the positions of the center of the receiving transducer and a current integration point on the defect surface. The polarization vector $\mathbf{d}^{(2)}$ is perpendicular to the direction vector $\mathbf{e}^{(2)}$ and is lying in the vertical plane of incidence. It may be calculated from:

$$\mathbf{d}^{(2)} = \frac{\mathbf{e}^{(2)} \times \mathbf{s}}{\left|\mathbf{e}^{(2)} \times \mathbf{s}\right|} \times \mathbf{e}^{(2)}, \tag{24}$$

where $\mathbf{s}$ is in the unit vector normal to the tested object surface (see Fig. 3). This way it is shown how to calculate the stress tensor components appearing in the reciprocity Eq. (14).

Now the particle velocity components $v_i^{(1)}$ appearing in Eq. (14) would be determined. These are the components of solution (1), which takes into account the interaction between the ultrasonic wave generated by the transmitting transducer and the embedded model defect. The exact solution to this type of problem is very difficult, even for simple model defects. To solve this problem in a simplified way, the Kirchhoff approximation would be applied. In the Kirchhoff approximation, the interaction of the ultrasonic beam with the material discontinuity is treated as the interaction of a (locally) plane wave with the plane boundary in the propagation medium. Such seminal problems have well-known solutions given by reflection and refraction coefficients (Auld, 1973; Schmerr, 2016). In the model presented in this paper, crack-like planar defects with transverse dimensions considerably larger than the ultrasonic wavelength are considered. According to an in-depth analysis of the Kirchhoff approximation given by Huang *et al.* (2006), it should give reasonably accurate results in modeling such types of defects. The calculation details of the vector $\mathbf{v}^{(1)}$ can be found in (Katz *et al.*, 2021). The most important steps in deriving the final formula are presented further.

The wave approaching to a certain small element dS on the defect surface can be characterized by its direction vector $\mathbf{e}_i^{(1)}$ unit polarization vector $\mathbf{d}_i^{(1)}$ and an amplitude of the particle velocity $v_i^{(1)}$. The coordinates of these vectors can be calculated from the positions of the center of the transmitting transducer and the center of the element dS on the defect surface. Then the incident wave is decomposed into two standard polarities considered in relation to the plane of the dS element, shear vertical (SV), and shear horizontal (SH), with polarization vectors of $\mathbf{d}_{i\mathrm{SV}}$ and $\mathbf{d}_{i\mathrm{SH}}$. It is important to note that $\mathbf{d}_{i\mathrm{SV}}$ and $\mathbf{d}_{i\mathrm{SH}}$ are not unit vectors but the mutually perpendicular vector components whose sum gives the unit directional vector of the incident wave:

$$\mathbf{d}_i^{(1)} = \mathbf{d}_{i\mathrm{SV}}^{(1)} + \mathbf{d}_{i\mathrm{SH}}^{(1)}. \tag{25}$$

First, let us consider the incident wave of SV-type. For such a type of wave, there are generally two reflected waves, SV- and L-type. The SV-wave reflects at an angle equal to the incidence angle, and its directional vector $\mathbf{e}_{r\mathrm{SV}}$ and polarization vector $\mathbf{d}_{r\mathrm{SV}}$ can be calculated from the knowledge of $\mathbf{e}_i$ and $\mathbf{d}_{i\mathrm{SV}}$. The reflection coefficient for this type of wave, $R^{(\mathrm{SV,SV})}$, can be calculated from the standard formula for plane waves, as in (Schmerr, 2016), based on the incidence angle on the defect surface. The L-type wave reflects at an angle that can be calculated from Snell's law. The polarization direction of that wave is parallel to the propagation direction $\mathbf{e}_{rL}$. The reflection coefficient $R^{(L,\mathrm{SV})}$ can be calculated from the standard formula. These three waves are summed up to give the first part of the solution (1) arising from the SV component of the incident wave:

$$\mathbf{v}_{\mathrm{SV}}^{(1)} = v^{(1)}\left(\mathbf{d}_{i\mathrm{SV}} + R^{(\mathrm{SV,SV})}\mathbf{d}_{r\mathrm{SV}} + R^{(P,\mathrm{SV})}\mathbf{d}_{rL}\right). \tag{26}$$

The vectors $\mathbf{d}_{r\mathrm{SV}}$ and $\mathbf{d}_{rL}$ are not of the unit length but have the same reduced length as the vector $\mathbf{d}_{i\mathrm{SV}}$.

Let us therefore consider the incident wave of the SH-type. For such an incident wave, there is only one reflected wave of the SH-type. The SH-wave reflects at an angle equal to the incidence angle and its polarization vectors are given by

$$\mathbf{d}_{r\mathrm{SH}} = \mathbf{d}_{i\mathrm{SH}}. \tag{27}$$

The vector $\mathbf{d}_{r\mathrm{SH}}$ is not of a unit length but has a reduced length, the same as the vector $\mathbf{d}_{i\mathrm{SH}}$. Based on the above, an expression for the second part of the solution (1) caused by the SH polarization of the incident wave can be written as

$$\mathbf{v}_{\mathrm{SH}}^{(1)} = v^{(1)}\left(\mathbf{d}_{i\mathrm{SH}} + R^{(\mathrm{SH,SH})}\mathbf{d}_{r\mathrm{SH}}\right). \tag{28}$$

Having calculated particle velocities for both polarizations of the incident wave, the total particle velocity of the ultrasonic vibration for the solution (1) can be written as

$$\mathbf{v}^{(1)} = \mathbf{v}_{\mathrm{SV}}^{(1)} + \mathbf{v}_{\mathrm{SH}}^{(1)} = v^{(1)}\left(\mathbf{d}^{(1)} + R^{(\mathrm{SV,SV})}\mathbf{d}_{r\mathrm{SV}} + R^{(\mathrm{SH,SH})}\mathbf{d}_{r\mathrm{SH}} + R^{(P,\mathrm{SV})}\mathbf{d}_{rL}\right). \tag{29}$$

To simplify the notation, all the vectors defining the polarizations and relative amplitudes of waves acting on the defect surface can be grouped into one vector $\mathbf{D}^{(1)}$:

$$\mathbf{D}^{(1)} = \mathbf{d}^{(1)} + R^{(\mathrm{SV,SV})}\mathbf{d}_{r\mathrm{SV}} + R^{(\mathrm{SH,SH})}\mathbf{d}_{r\mathrm{SH}} + R^{(P,\mathrm{SV})}\mathbf{d}_{rL}. \tag{30}$$

Similarly, the expression for $\mathbf{v}^{(1)}$ can be rewritten as

$$\mathbf{v}^{(1)} = v^{(1)}(\mathbf{x},\omega)\,\mathbf{D}^{(1)}. \tag{31}$$

At this stage we have determined all the elements necessary for the reciprocity Eq. (14), which may be rewritten in a more straightforward form where we also neglected all constants before the integral sign:

$$V_R(\omega) = \mathrm{SF}(\omega)\int\limits_{S_d} v^{(1)}v^{(2)}\left[\left(\mathbf{D}^{(1)}\cdot\mathbf{d}^{(2)}\right)\left(\mathbf{e}^{(2)}\cdot\mathbf{n}\right) + \left(\mathbf{D}^{(1)}\cdot\mathbf{e}^{(2)}\right)\left(\mathbf{d}^{(2)}\cdot\mathbf{n}\right)\right]\mathrm{dS}, \tag{32}$$

where $v^{(1)}(\mathbf{x},\omega)$ is the amplitude of solution (1) calculated from Eq. (11) for transmitting transducer at position $(0,0,0)$, $v^{(2)}(\mathbf{x},\omega)$ is the amplitude of solution (2) calculated from Eq. (11) for receiving transducer at position $(\Delta X,0,0)$, $\mathbf{D}^{(1)}$ is the polarization vector for the solution (1), $\mathbf{d}^{(2)}$ is the polarization vector for the solution (2).

Integral Eq. (32) provides a workable solution for numerical calculations of amplitudes of ultrasonic echoes reflected from the planar model defects of any size and orientation. The main difference of this solution in relation to the known solutions obtained within the quasi-static ultrasonic model is that all the quantities related to solution (2) are calculated for the receiving position of the transducer, which is shifted by ΔX from the initial position where the ultrasonic transmission takes place. Equation (32) allows for the computation of receiving

voltage at a fixed frequency ω. To calculate the signal waveform in the time domain, it is necessary to calculate $V_R(\omega_k)$ for the series of frequencies contained in the SF spectrum and then calculate the inverse Fourier transform using the FFT algorithm. Then, the maximum echo amplitude in the time domain can be determined.

6. Calculation of ultrasonic field generated by angle beam probes

The presented model allows for computer simulation of several important aspects of ultrasonic inspection of railway rails. The first basic functionality of the software is the calculation of ultrasonic fields generated in the tested material by ultrasonic probes with different parameters (frequency, refraction angle, transducer size, bandwidth, etc.). To illustrate this functionality, we calculated the ultrasonic fields of three angle beam probes that could be used for detecting transversal defects in the rail head. There were angle beam shear wave probes: T60°, T70°, and T80° of the same central frequency (f_0 = 2 MHz), bandwidth (BW = 50 %), and transducer size (14 mm × 14 mm). The refraction wedges of all probes were made of PMMA. The probes can be positioned on the rail head center, and their beams are directed along the plane of symmetry of the rail (X-Y plane).

In Fig. 4, the beam cross-sections in the X-Y plane are illustrated with the color-coded maps of the particle velocity amplitude. It should be emphasized that there are distributions of maximum amplitudes of short pulses generated by the typical ultrasonic probes, not the amplitudes of a monochromatic wave of a single

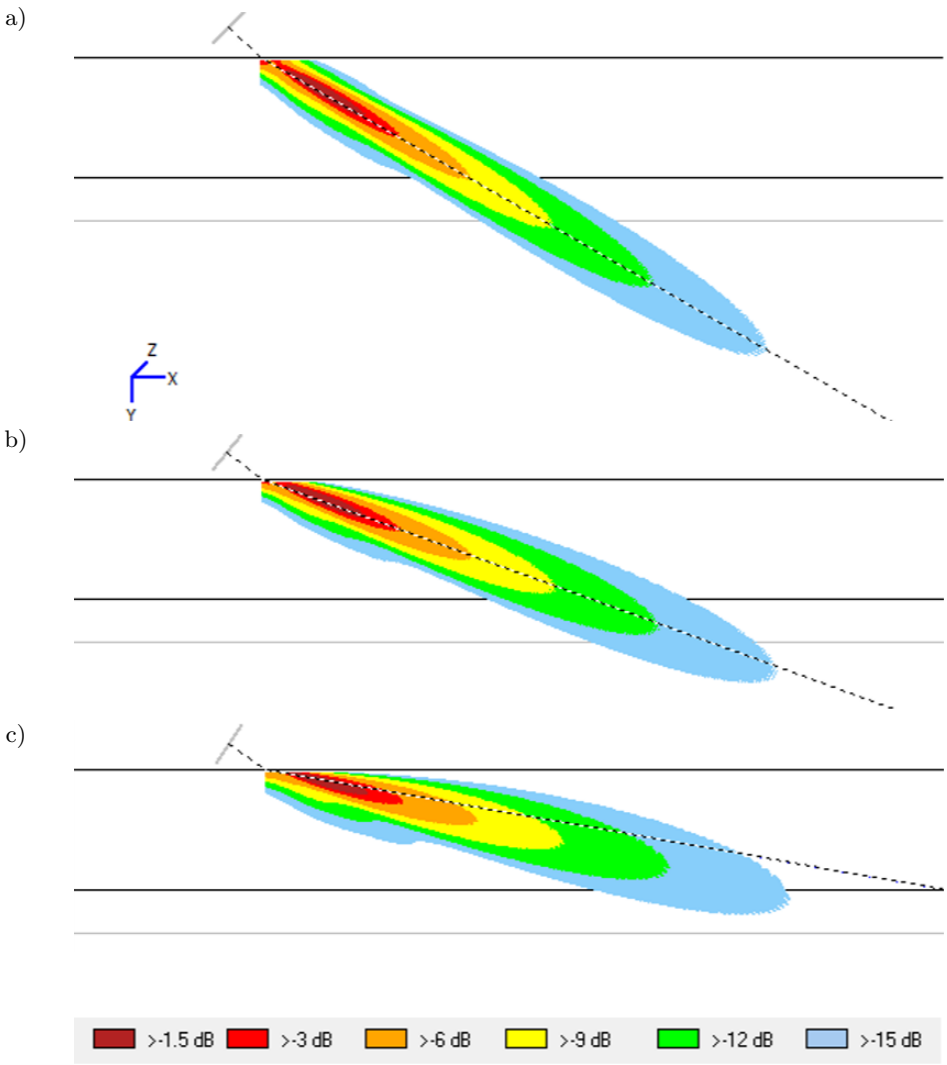

Fig. 4. Ultrasonic beam cross-sections for three similar angle beam probes with different refraction angles: a) 2T60-A14x14, b) 2T70-A14x14, c) 2T80-A14x14 (the probe codes are explained in the text).

frequency. All field distributions are normalized to the maximum value in the field of a given probe, so the maps do not illustrate correctly the differences in absolute amplitudes between different probes. Actually, the highest field amplitude is observed for the near-field maximum of the T60° probe. For the T70° probe, the maximum is only slightly smaller (0.2 dB), but for the T80° probe, it is reduced by 2.3 dB in comparison to the T60° probe.

Another interesting observation is the considerable deviation of the acoustic axis from the geometric axis (the dashed line on the field graphs) for the T80° probe. This is a manifestation of the fact that the transmission coefficient of ultrasonic waves on the wedge-material border quickly drops near the 2nd critical angle. Based on presented distributions of ultrasonic fields, one can draw the conclusion that using angle beam probes with nominal refraction angles higher than 70° to 75° would not be effective for detecting rail head transversal defects despite the fact that their nominal refraction angle is better suited to detecting vertical defects.

Figure 5 illustrates the ultrasonic beam cross sections for three T70° probes, distinguished solely by their frequency BW: a narrow band probe (BW = 20 %), a medium band probe (BW = 50 %), and a broad band probe (BW = 90 %).

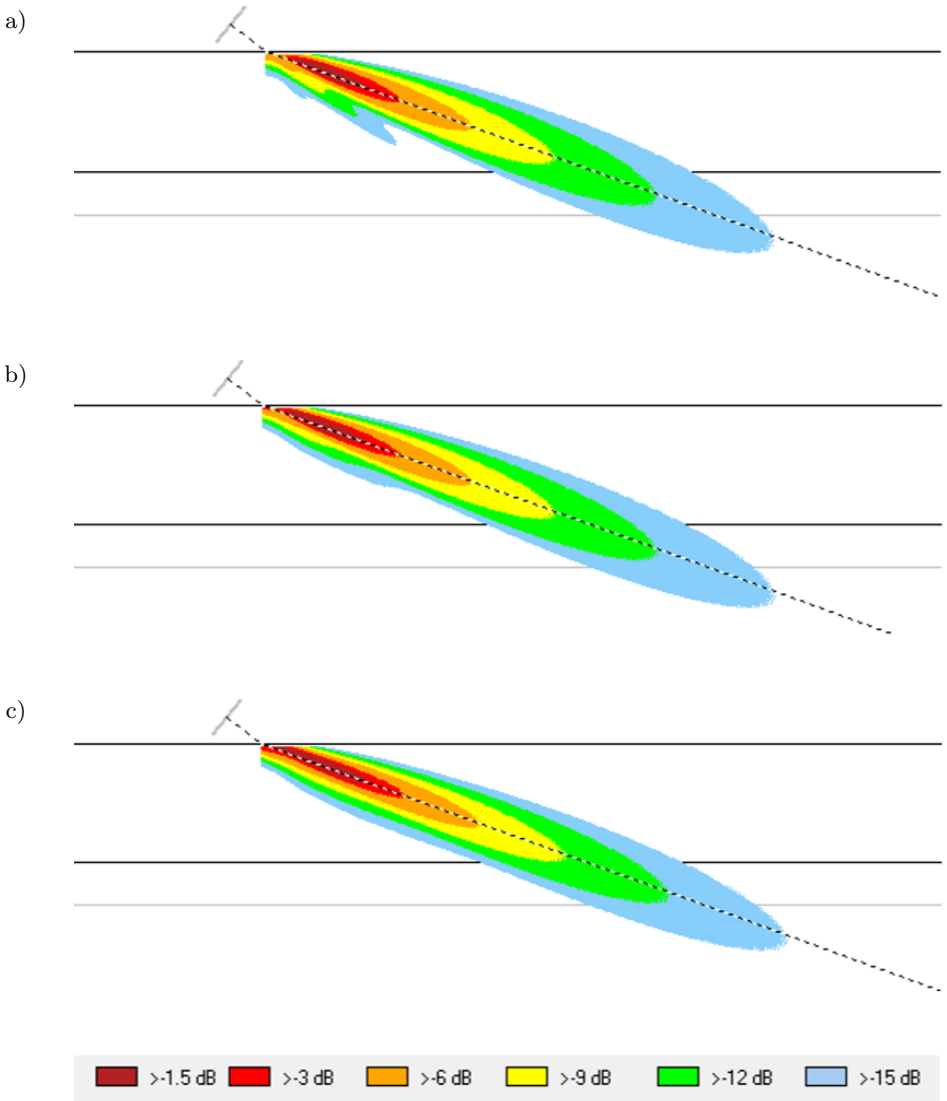

Fig. 5. Ultrasonic beam cross-sections for three 2T70-A14x14 type probe differing only in frequency BW: a) 20 %, b) 50 %, c) 90 %.

Overall, the ultrasonic fields for all three probes are similar, with only one noticeable difference. The field for the narrow band probe, as shown in picture Fig. 5a, displays one side lobe below the main lobe. The ultrasonic

fields of the other two probes are more uniform and do not exhibit any side lobes. The last example of using the simulation program for calculation of ultrasonic beams of angle beam probes concerns the influence of the attenuation coefficient of the rail material on the distribution of ultrasonic field. As mentioned in Sec. 3 the attenuation coefficient of ultrasonic waves in rail steel was taken from the tables given in (Ono, 2020a; 2020b). These attenuation coefficient values are used as standard in our simulation program for calculation of ultrasonic field distributions. We have checked whether assuming a zero attenuation coefficient in railway steel would significantly affect the calculated ultrasonic fields.

In Fig. 6 there are presented the ultrasonic beams of 2T70-A14x14 probe (BW = 50 %) calculated with the attenuation coefficient taken from (Ono, 2020b) and the attenuation coefficient assumed to be zero.

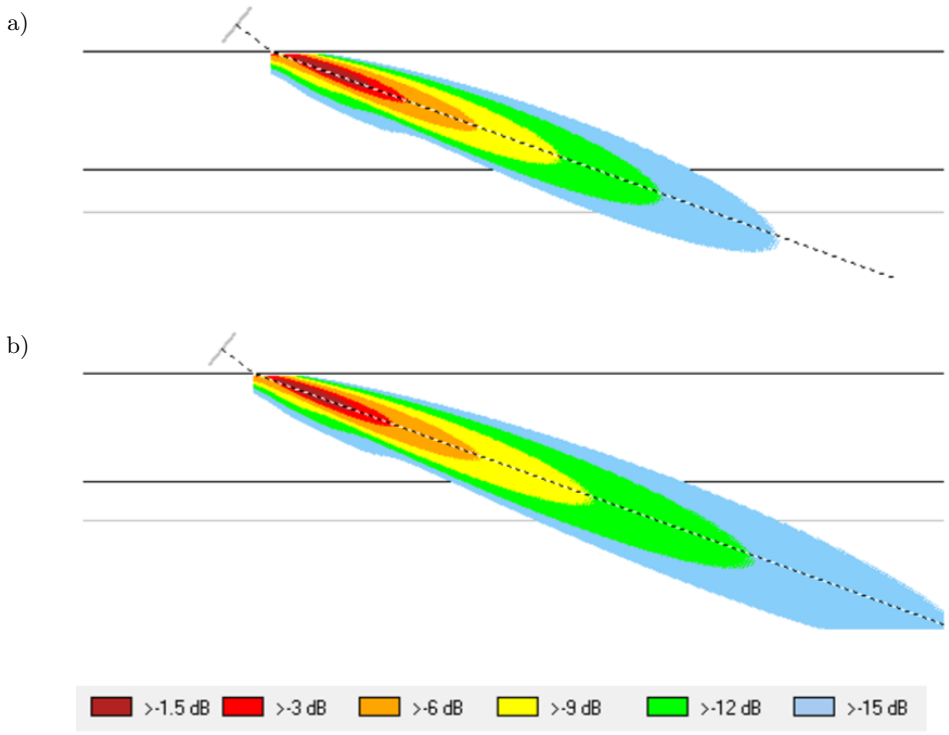

Fig. 6. Ultrasonic beam cross-sections for 2T70-A14x14 probe calculated with assumption of attenuation coefficient: a) taken from (Ono, 2020b), b) equal to zero.

As can be seen, the attenuation coefficient has a considerable effect on the ultrasonic field distribution of a typical 2 MHz shear wave probe commonly used in rail inspections. It means that the unjustified assumption of a zero attenuation coefficient of railway steel would lead to significant overestimation of the simulated field and misleading conclusions concerning the probe sensitivity zone.

7. Simulation of ultrasonic echo envelopes during rail scanning

The second fundamental feature of the prepared software SymUT is the calculation of ultrasonic echo amplitudes from simple model defects (circles, rectangles) of various sizes, orientations, and positions. Considering the operating principles of automated rail testing systems, the most significant outcome of ultrasonic inspections is the echo envelope recorded for each testing probe while it moves along the rail over defect locations. Based on the amplitudes and shapes of the recorded envelopes, the defects are classified and displayed on the B-scan diagrams with the appropriate colors.

The software allows for the calculation of echo envelopes for defined ultrasonic probes moving along the rail axis over the simulated model defects. It calculates the echo amplitudes for the successive positions of the probe along the X-axis in the points, which are determined by the scanning speed and repetition frequency of

the testing system. In this way, the computer simulation closely imitates the operation of a real rail inspection system, taking into account its operating parameters.

We simulated echo envelopes for two different defects, one located in the rail head and the other in the rail foot. The main goal of the performed simulations was to show the differences between our new dynamic ultrasonic testing model and the standard quasi-static model. As the first example, we calculated the echo envelopes for the circular defect situated in the central part of the rail head which is supposed to simulate a common transversal defect of the rail head called a tache ovale (Kumar, 2006). The center of the model defect was assumed at Y_d = 15 mm and its diameter D_d was 10 mm. It was assumed that the defect plane is deflected from the rail transversal plane by 10° towards the bottom of the rail (see Fig. 7).

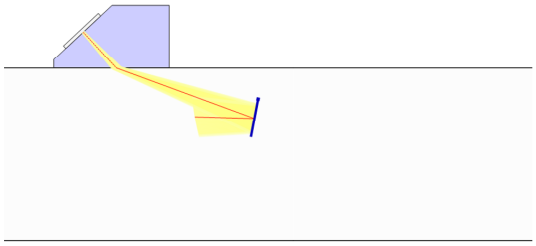

Fig. 7. Simulated probe (rail) defect configuration with circular defect of diameter D_d = 10 mm deviated by θ = 10° from the rail transversal plane.

The model defect was detected with the 2T70-A14x14 ultrasonic probe described in Sec. 6. The probe was guided along the center of the rail head with its beam directed along the rail, parallel to the scanning direction. The simulated test configuration is illustrated in Fig. 7.

As can be readily seen from Fig. 7, in the simple geometric approximation of ray tracing, the implemented defect would not be detected due to unfavorable orientation with respect to the incident beam. However, in the model presented in the following paper, which takes into account the diffraction phenomena, such defects can be detected with a reasonable echo amplitude.

In Fig. 8, the simulated pulse echo envelope calculated for the above testing configuration can be seen. The following assumptions were taken: the scanning speed: 80 km/h and the repetition frequency of the ultrasonic system: 5 kHz. The black horizontal line on the chart shows the reference level representing the maximum amplitude of the pulse echo from the standard side-drilled hole (SDH) reflector of 6 mm diameter situated 20 mm below the rail surface. The echo amplitude from such a cylindrical reflector was calculated using the model of (Lopez-Sanchez *et al.*, 2005), because the model discussed in this paper was developed only for flat reflectors. This mixed approach implies that the test sensitivity of the ultrasonic system should be set up when the inspection wagon is stationary or is moving over the reference rail with a low speed. The triangles in Fig. 8 represent the points of the dynamic echo envelope, which would be registered by the ultrasonic system when scanning the rail at a speed of

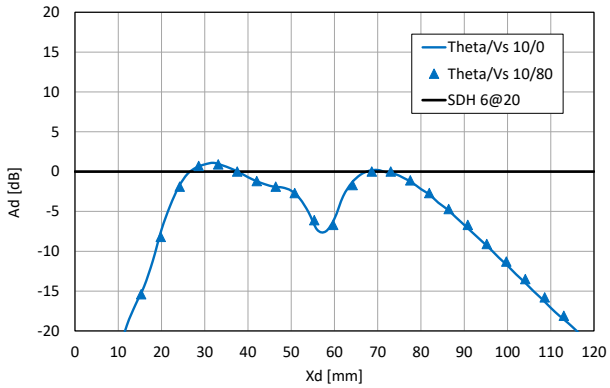

Fig. 8. Simulated echo envelope of the model defect (D = 10 mm, θ = 10°) detected with 2T70-A14x14 probe scanning the rail with speed of 80 km/h.

80 km/h and a repetition rate of 5 kHz. The scanning direction is assumed to be parallel to the viewing direction of the ultrasonic probe. For comparison, a continuous line shows the echo envelope calculated using the quasi-static ultrasonic model, which neglects the movement of the ultrasonic probe during the transmission-reception cycle.

As can be seen, the main difference between the dynamic and the quasi-static model is the discretization of the echo envelope due to the fact that a 5 kHz system running with a speed of 80 km/h fires ultrasonic pulses every 4.4 mm along the rail length. The difference in amplitude for the corresponding points of both envelopes does not exceed 0.3 dB. It seems that a scanning speed of 80 km/h does not create major problems for the analyzed testing configuration. Therefore, in the next simulation, the scanning speed was increased to 160 km/h. The run of obtained echo envelopes is shown in Fig. 9. In this case, echo envelopes were calculated for both scanning directions, i.e., parallel and antiparallel to the viewing direction of the ultrasonic probe. In Fig. 9 the black horizontal line on the chart showing the reference level representing the amplitude of the pulse echo from the standard SDH reflector is also presented. With higher testing speed, the problem of sparse sampling (in this case, every 8.9 mm) becomes more evident. It consists of the fact that a high echo amplitude from a given defect is recorded only at a few points, which may lead to uncertainty in the interpretation of the indications. This is because algorithms of automatic defect detection during rail scanning require several (often 5–6) consecutive echo registration cycles of signal level situated above the recording level to qualify the indication as significant. It means that the algorithms of indication discrimination should be carefully adjusted to the scanning speed based on the results of similar simulations.

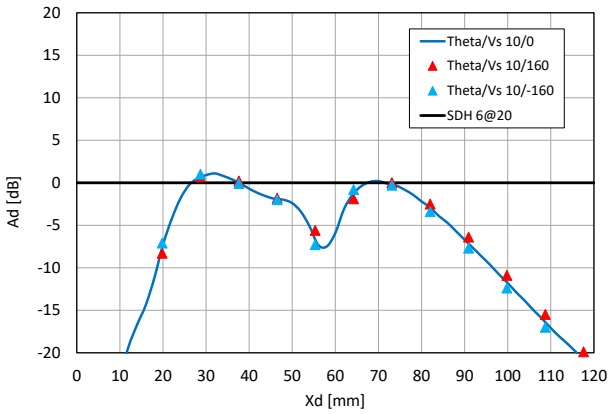

FIG. 9. Simulated echo envelopes of the model defect (D = 10 mm, θ = 10°) obtained with T70-14x14 probe scanning the rail with speed of 160 km/h in two directions, i.e., parallel (red triangles) and antiparallel (blue triangles), to the viewing direction of the ultrasonic probe.

Another conclusion from the graphs presented in Fig. 9 is that the difference between the amplitudes of the envelopes calculated from the dynamic and quasi-static models increased to ±0.7 dB. This difference is solely due to the fact that the ultrasonic probe moves during the transmitting-receiving cycle, and its sign depends on whether the movement is parallel or antiparallel to the viewing direction of the probe.

8. Conclusions

In the paper, the new theoretical model of ultrasonic testing with high scanning speed was presented. Unlike the standard quasi-static models used so far, it does not assume that the scanning speed of the ultrasonic probe is negligible in comparison to the speed of ultrasonic waves in the tested material. In consequence, it takes into account that the ultrasonic probe changes its position during the transmitting-receiving cycle of the ultrasonic system. The new model is based on well-established principles of ultrasonic theory, such as the Rayleigh–Sommerfeld integral, reciprocity relation, and the Kirchhoff approximation, but implements a new concept of variable probe position at the moments of pulse transmission and reception. The ΔX shift in the probe position is calculated based on the defect position relative to the probe at the pulse firing time, the probe scanning speed, and the speed of sound in the tested material.

The model allows for the calculation of the ultrasonic field generated by the probe in the tested material and the amplitudes of the ultrasonic echoes reflected from the simple model defects implemented in the tested material. Based on the theoretical model, the SymUT software was developed and designed specifically for the simulation of high-speed ultrasonic testing of railway rails in track. Thanks to the semi-analytical nature of the developed model, the program can be executed on standard PC computers with reasonably short computational times. It enables effective computer simulation of many test configurations and optimization of ultrasonic probes for detection of different types of defects occurring in railway rails.

In the last part of the paper, example simulations were presented and discussed in the context of increased speed of automated testing of rails. The simulations performed showed that the defect echo envelopes calculated from the quasi-static and dynamic models differ in two ways. Firstly, the echo envelopes calculated within the dynamic model are discrete, which is a direct consequence of the limited repetition frequency (c.a. 5 kHz) of the ultrasonic system. The sparse sampling along the rail length can reduce the detectability of some defects and requires modification of automated indication discrimination algorithms. On the other hand, the differences in echo amplitudes in corresponding points of envelopes calculated according to dynamic and quasi-static models are not too big and reach 3 dB for a scanning speed of 160 km/h. In general, it can be stated that there are no fundamental physical obstacles to increasing the speed of automatic testing of railway rails to 160 km/h, assuming that the technical problems related to the mechanical guidance of the probes and their acoustic coupling are solved.

Fundings

This work was supported by the project no. BRIK2/0013/2022 of the Polish National Centre for Research and Development.

Conflict of interests

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Authors' contribution

Sławomir Mackiewicz conceptualized the study and wrote the original draft. Zbigniew Ranachowski performed the required measurements and delivered the data analysis. Tomasz Katz created the graphs and took part in the measurements. Tomasz Dębowski prepared the instrumentation to verify the presented results. Grzegorz Starzyński contributed to data interpretation. All authors reviewed and approved the final manuscript.

Acknowledgments

The paper was written as a result of the implementation of the Project no. BRIK2/0013/2022, 'A joint undertaking of the National Centre for Research and Development – PKP Polskie Linie Kolejowe S.A. to support scientific research and development work in the area of railway infrastructure, entitled Research and Development in Railway Infrastructure BRIK II.'

References

1. Auld B.A. (1973), *Acoustic Fields and Waves in Solids*, p. 167, John Wiley & Sons.
2. Auld B.A. (1979), General electromechanical reciprocity relations applied to the calculation of elastic waves scattering coefficients, *Wave Motion*, **1**(1): 3–10, https://doi.org/10.1016/0165-2125(79)90020-9.
3. Calmon P., Lhémery A., Lecœur-Taïbi I., Raillon R., Paradis L. (1998), Models for the computation of ultrasonic fields and their interaction with defects in realistic NDT configurations, *Nuclear Engineering and Design*, **180**(3): 271–283, https://doi.org/10.1016/S0029-5493(97)00299-9.

4. Darmon M., Chatillon S. (2013), Main features of a complete ultrasonic measurement model: Formal aspects of modeling of both transducers radiation and ultrasonic flaws responses, *Open Journal of Acoustics*, **3**(3A): 43–53, https://doi.org/10.4236/oja.2013.33A008.
5. Deschamps G.A. (1972), Ray techniques in electromagnetics, *Proceedings of the IEEE*, **60**(9): 1022–1035, https://doi.org/10.1109/PROC.1972.8850.
6. European Committee for Standardization (2016), *Railway applications – Infrastructure – Non-destructive testing on rails in track. Part 1: Requirements for ultrasonic inspection and evaluation principles* (European Standard EN 16729-1:2016), CEN, Brussels.
7. European Committee for Standardization (2018), *Railway applications – Infrastructure – Non-destructive testing on rails in track. Part 3: Requirements for identifying internal and surface rail defects* (European Standard EN 16729-3:2018), CEN, Brussels.
8. Gengembre N., Lhemery A. (2000), Pencil method in elastodynamics: Application to ultrasonic field computation, *Ultrasonics*, **38**(1–8): 495–499, https://doi.org/10.1016/s0041-624x(99)00068-2.
9. Heckel T., Casperson R., Rühe S., Mook G. (2018), Signal processing for non-destructive testing of railway tracks, [in:] *Proceedings of 44th Annual Review of Progress in Quantitative Nondestructive Evaluation*, **1949**(1): 030005, https://doi.org/10.1063/1.5031528.
10. Huang R., Schmerr L.W., Sedov A., Gray T.A. (2006), Kirchoff approximation revisited – Some new results for scattering in isotropic and anisotropic elastic solids, *Research in Nondestructive Evaluation*, **17**(3): 137–160, https://doi.org/10.1080/09349840600787956.
11. International Organization for Standardization (2020), *Non-destructive testing – Characterization and verification of ultrasonic test equipment. Part 2: Probes* (ISO Standard No. 2232-2:2020), https://www.iso.org/obp/ui/es/#iso:std:iso:22232:-2:ed-1:v1:en.
12. International Union of Railway (2002), *Rail defects* (UIC Code 712-R), 4th ed., https://www.scribd.com/document/338561091/UIC-Code-712-R-2002-Rail-Defects (access: 28.10.2025).
13. Katz T., Mackiewicz S., Ranachowski Z., Kowalewski Z.L., Antolik Ł. (2021), Ultrasonic detection of transversal cracks in rail heads – Theoretical approach, *Engineering Transactions*, **69**(4): 437–456, https://doi.org/10.24423/EngTrans.1695.20211220.
14. Kumar S. (2006), *Study of rail breaks: Associated risks and maintenance strategies*, Technical report, Division of Operation and Maintenance Engineering, Luleå Railway Research Center (JVCT), Luleå University of Technology, https://www.diva-portal.org/smash/get/diva2:995250/FULLTEXT01.pdf (access: 28.10.2025).
15. Lopez-Sanchez A.L., Kim H.-J., Schmerr L.W.Jr., Sedov A. (2005), Measurement models and scattering models for predicting the ultrasonic pulse-echo response from side-drilled holes, *Journal of Nondestructive Evaluation*, **24**(3): 83–96, https://doi.org/10.1007/s10921-005-7658-4.
16. Mackiewicz S., Ranachowski Z., Katz T., Dębowski T., Starzyński G., Ranachowski P. (2024), Modeling of acoustic coupling of ultrasonic probes for high-speed rail track inspection, *Archives of Acoustics*, **49**(2): 255–266, https://doi.org/10.24425/aoa.2024.148787.
17. Ono K. (2020a), A comprehensive report on ultrasonic attenuation of engineering materials, including metals, ceramics, polymers, fiber-reinforced composites, wood and rocks, *Applied Sciences,* **10**(7): 2230, https://doi.org/10.3390/app10072230.
18. Ono K. (2020b), Dynamic viscosity and transverse ultrasonic attenuation of engineering materials, *Applied Sciences*, **10**(15): 5265, https://doi.org/10.3390/app10155265.
19. Raillon R., Lecoeur-Taïbi I. (2000), Transient elastodynamic model for beam defect interaction: Application to nondestructive testing, *Ultrasonics*, **38**(1–8): 527–530, https://doi.org/10.1016/S0041-624X(99)00067-0.
20. Schmerr L.W.Jr. (2016), *Fundamentals of Ultrasonic Nondestructive Evaluation. A modelling approach*, 2nd ed., Springer Nature, Basel, https://doi.org/10.1007/978-3-319-30463-2.
21. Thomas H.-M., Heckel T., Hanspach G. (2007), Advantage of a combined ultrasonic and eddy current examination for railway inspection trains, *Insight*, **49**(6): 341–344, http://doi.org/10.1784/insi.2007.49.6.341.
22. Zumpano G., Meo M. (2006), A new damage detection technique based on wave propagation for rails, *International Journal of Solids and Structures*, **43**(5): 1023–1046, https://doi.org/10.1016/j.ijsolstr.2005.05.006.

Archives of Acoustics Vol. 51, No. 2, pp. 377–385 (2026), https://doi.org/10.24423/archacoust.2026.4351

Technical Note

Semi-Supervised Learning for Sediment Classification Using Convolutional Neural Networks with Digital Elevation Model and Backscatter Data

Paweł Nadachowski[(1)], Zbigniew Łubniewski[(1)*], Karolina Trzcińska[(2)], Radosław Wróblewski[(3)], Maria Rucińska[(2)], Jarosław Tęgowski[(2)]

[(1)] *Faculty of Electronics, Telecommunications and Informatics, Department of Geoinformatics, Gdansk University of Technology*
Gdansk, Poland

[(2)] *Faculty of Oceanography and Geography, Department of Geophysics, University of Gdansk*
Gdynia, Poland

[(3)] *Maritime Institute, Gdynia Maritime University*
Gdansk, Poland

*Corresponding Author: lubniew@eti.pg.edu.pl

Received October 2, 2025; accepted February 3, 2026;
version of record June 3, 2026; published issue June 24, 2026.

Accurate sediment classification is crucial for advancing marine research, environmental monitoring, and sustainable seabed use. However, acquiring large amounts of labeled data in such settings is often challenging, expensive, and time-consuming. To address this limitation, a semi-supervised learning framework has been proposed that leverages convolutional neural networks for sediment classification using both labeled and unlabeled data. The approach utilizes pseudo-labeling, where confident predictions on unlabeled samples are iteratively incorporated into training to enhance model generalization. The method is applied and evaluated on a dataset that includes multi-modal inputs such as a digital elevation model and multibeam sonar backscatter data. Experimental results indicate that semi-supervised learning with convolutional neural networks can achieve high classification accuracy in scenarios characterized by limited labeled data and a large volume of unlabeled data. This approach highlights the potential of deep learning combined with semi-supervised strategies for efficient underwater environment classification.

Keywords: convolutional neural network (CNN), deep learning, digital elevation model (DEM), multibeam sonar backscatter, pseudo label, residual neural network (ResNet), sediment classification, semi-supervised learning (SSL).

1. Introduction

Supervised seabed sediment classification requires a large amount of ground truth data, which is very expensive to obtain. This is especially troublesome if there is a need to use deep learning methods, such as convolutional neural networks (CNNs) (Krizhevsky *et al.*, 2012), which typically require large-scale annotated datasets to achieve optimal generalization. In marine and geological environments, collecting accurately labeled data is particularly challenging due to logistical constraints, high costs of field surveys, and the need for expert interpretation of sediment types. Consequently, there is a growing interest in exploring data-efficient learning strategies to reduce dependency on large amounts of ground truth data (Qin *et al.*, 2021; Zhao *et al.*, 2023).

To alleviate this problem, this study examines how semi-supervised learning (SSL) (Hady, Schwenker, 2013) methods can improve model performance by leveraging both labeled and unlabeled data. One such SSL technique is the pseudo-labeling (Lee, 2013) method, which offers a simple yet effective approach to extend the training dataset. In pseudo-labeling, a supervised model initially trained on a small subset of labeled data is used to predict labels for the unlabeled samples. Predictions with high confidence are then treated as ground truth (pseudo-labels) and used for further training, effectively augmenting the training dataset with automatically annotated examples.

By incorporating pseudo-labeled samples, CNNs can be trained to better generalize from sparse labeled data, capturing more complex patterns and improving classification accuracy on unseen samples. This method is especially relevant in sediment classification tasks, where expert-labeled examples are scarce and expensive to obtain. The pseudo-labeling framework also allows for the use of different confidence thresholds to control the quality of the generated pseudo-labels. This study evaluates the impact of different confidence thresholds on model performance, aiming to identify the optimal balance between label quality and data coverage. Results demonstrate that, when applied carefully, pseudo-labeling can achieve high classification accuracy while minimizing the need for exhaustive manual annotation.

2. Study site and methods

The experiments were conducted at a single study site, where multiple processing and analysis steps were carried out using different techniques. Each step was aimed at optimizing the classification process and improving overall performance. The applied approach, which included data acquisition, ground-truth labeling, model selection and SSL implementation, ultimately led to high classification accuracy.

2.1. Study site

The study area encompasses a section of the seabed, covering nearly 10 km, at depths ranging from 3.5 m to 22 m. It is located less than one kilometer offshore, north of the village of Rowy, in the southern Baltic Sea, in the Polish Territorial Sea, within the Natura 2000 site (code PLB990002, Birds Directive), partially overlapping with the Słowiński National Park. The study site location is presented in Fig. 1.

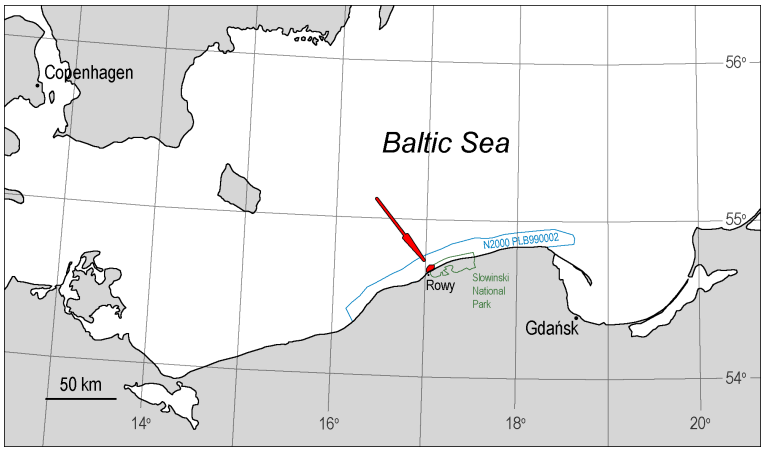

Fig. 1. Study site location.

The seabed structure and morphology in this region are primarily shaped by glacial processes that occurred during the Pleistocene, followed by deglaciation and subsequent developments associated with the evolution of the Baltic Sea during the Holocene. The seabed sediment composition includes Pleistocene glacial and fluvioglacial sediments, as well as Holocene marine deposits.

The predominant geomorphological feature of the area is a rugged relief formed by irregular, elevated outcrops of cohesive sediments, often with surface cobbles and boulders, interspersed with accumulations of sand and gravel.

The eastern and southeastern portions of the area are characterized by relatively level, gently undulating sandy surfaces. Similar sand accumulations are also present in patches across the central and western parts of the study area.

2.2. Data acquisition

Bathymetric and backscatter data were recorded between 2018 and 2020 using a Teledyne Reson SeaBat 7125 multibeam echosounder (MBES) mounted on the R/V Oceanograf vessel belonging to the University of Gdańsk. The MBES recording was performed at a frequency of 400 kHz using 512 beams with a width of 1° along-track and 0.5° across-track. Registration was performed in the QPS Qinsy software, combining information from MBES, sound velocity profiles in the water column, locations, and rotations of the measuring unit. A Global Navigation Satellite System/Inertial Navigation System supported by real time kinematic (RTK) corrections from the NAWGEO network (ASG-EUPOS[1]) provided real-time positioning with a few cm accuracy. A Valeport miniSVP Sound Velocity Profiler was used to record the speed of sound in the water column. The Octans gyrocompass and motion sensor were used to record the roll, pitch, heave, and heading of the vessel during measurements.

2.3. Ground-truth labeling

In the study area, ground-truth samples were obtained by collecting surface sediments using a Van Veen grab sampler and by recording video of the seabed with an underwater camera. The locations of the samples were recorded using GPS during measurement campaigns in 2018–2020. The sample locations were selected to take into account the diversity of sediments based on knowledge of the study area, bathymetric and backscatter maps, while maintaining, to the extent possible, an even spatial distribution of samples across the mapped area. However, due to the limited survey duration, the data set is not very large and contains 73 samples. The sediment samples were analyzed using granulometric methods in the laboratory, comprising Folk and Ward parameters as well as the Wentworth classification system. The underwater camera recordings were analyzed by an expert who assessed the type of sediment at the bottom, which was particularly important in places where it was not possible to collect samples with a grab, e.g., boulder fields. After analyzing all samples and recordings in the study area, generalizations were made and the sediments were divided into four classes: boulders (B), sandy gravel or gravelly sand (SG-GS), sand (S), very fine sand (VFS). Samples of recordings for each sediment type are presented in Fig. 2 and the visualization of bathymetry and backscatter data with the locations of different sediment types is shown in Fig. 3.

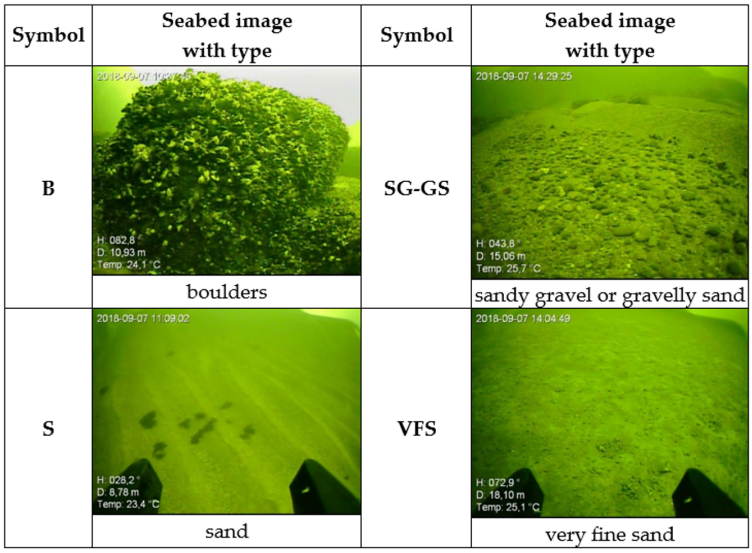

Fig. 2. Samples of recordings for each sediment type.

[1] https://www.asgeupos.pl

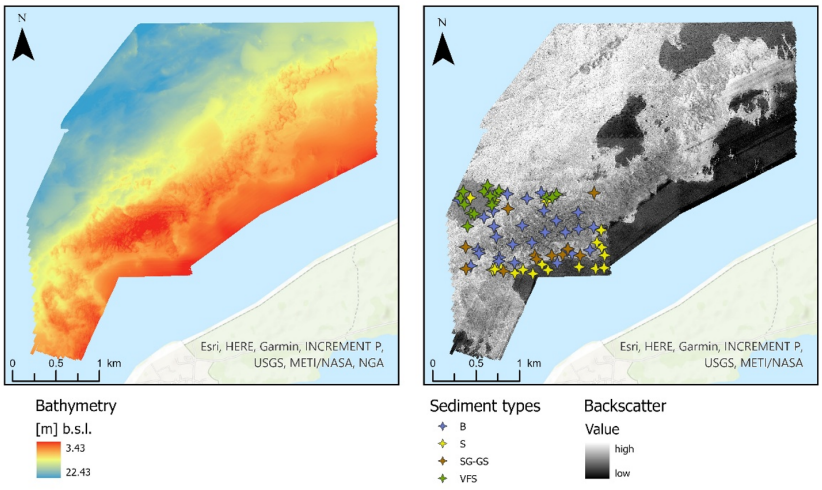

Fig. 3. Visualization of bathymetry and backscatter data with sediments.

2.4. Data preprocessing

Following the collection of bathymetric and backscatter data, it was processed to be ready for analysis. The MBES datasets were processed in QPS Qimera and Fledermaus Geocoder Toolbox (FMGT). In Qimera software, bathymetric data was cleaned and mosaicked, then exported to Geotiff with a resolution of 0.5 m × 0.5 m pixel size in the Universal Transverse Mercator (zone 33N) projected coordinate system. Similarly, the backscatter mosaic exported from FMGT was saved after manual removing of outliers as Geotiff. The backscatter mosaicking method used beam time series (snippets) and angle varying gain (AVG) correction within the Geocoder engine (Fonseca, Calder, 2005). The AVG method, commonly used to correct MBES angular dependency and normalize seafloor backscatter, was applied with default settings: 'flat' mode for noise reduction, 'blend' mosaicking for managing swath overlaps, and a 300-ping processing window (Schimel *et al.*, 2018; Parnum, Gavrilov, 2011).

Since convolutional neural networks were employed in this study, the preprocessed data was converted into a structured grid format suitable for the model input. Both labeled and unlabeled data were prepared in this way. Regarding the labeled dataset, a 16 × 16 grid was generated for each ground-truth point, capturing neighboring values from both the DEM and multibeam sonar backscatter channels. Next, the datasets were divided into two groups based on the year of acquisition: 31 samples from 2018 and 42 samples from 2019. In each group the data were imbalanced, with some classes significantly underrepresented. To address this issue, the imbalanced-learn (Lemaître *et al.*, 2017) library was used to perform oversampling of the minority classes to match the size of the majority class.

For the unlabeled dataset, 3000 random points were generated using QGIS tools, ensuring that they did not overlap with the area of the labeled data. For each of these points, a 16 × 16 grid was also created, incorporating neighboring values from both the DEM and multibeam sonar backscatter channels.

2.5. Convolutional neural network

CNNs are well-suited for solving problems in the field of computer vision. In this study, the input data consist of gridded representations of the digital elevation model (DEM) and multibeam sonar backscatter values, which can be treated as image-like inputs. Therefore, CNN-based architectures are a natural choice for addressing this classification task.

A specific type of convolutional network is the residual network (ResNet) (He *et al.*, 2016) which was employed in this study. ResNet architecture is known for its use of residual connections, which help mitigate the vanishing gradient problem and enable the effective training of deep neural networks. ResNet architecture has many variants and ResNet-18 was selected for this research. This particular architecture consists of 18 layers (17 convolutional and 1 fully connected) and is relatively lightweight, which makes it more suitable for training

on datasets with a limited number of labeled (ground-truth) samples. Its reduced complexity helps prevent overfitting while maintaining sufficient depth to capture spatial patterns in the input data. Visualization of ResNet-18 architecture is presented in Fig. 4.

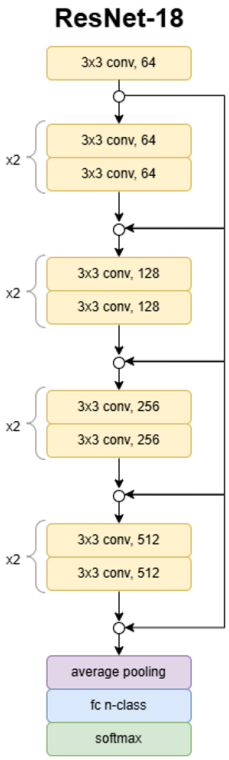

Fig. 4. Visualization of the ResNet-18 architecture.

2.6. Semi-supervised learning

The SSL is a technique that can be used when a small labeled dataset is available alongside a large volume of unlabeled data. A diagram of the required steps for the SSL is presented in Fig. 5.

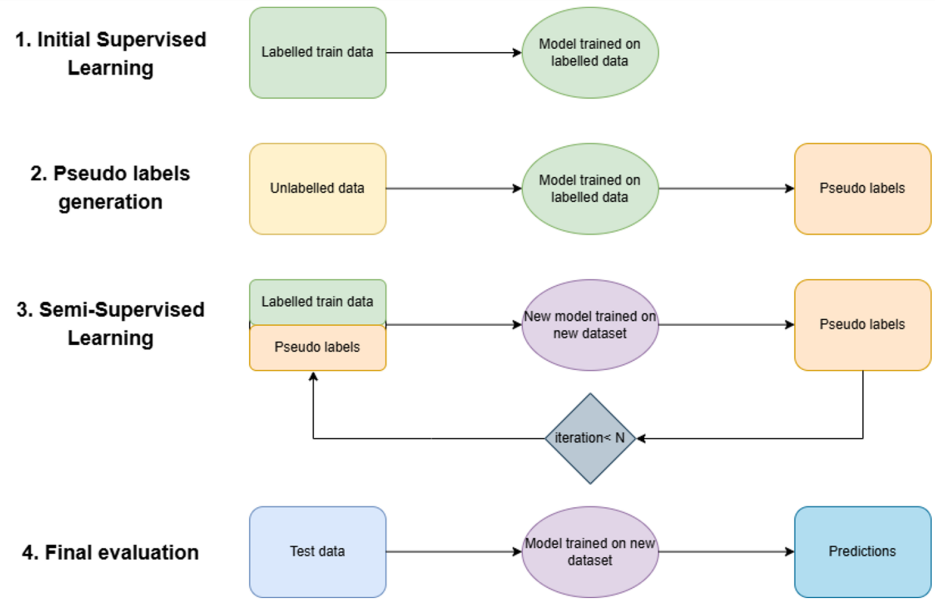

Fig. 5. Schema of required steps for SSL.

The first step is to train a ResNet-18 instance on a small labeled dataset. The labeled data was divided into training and test datasets based on the year of acquisition: samples from 2018 were used for training, and samples from 2019 for testing. In this study, the ResNet-18 model from the PyTorch (Paszke *et al.*, 2019) library's torchvision module was used. The hyperparameters were selected using a grid search method. The final values were: a learning rate of 1×10^{-7}, a batch size of 8, the CrossEntropyLoss as a loss function, and the Adam (Kingma, Ba, 2017) as an optimization algorithm.

After training the model on the supervised data, it can be used in the second step, where pseudo-label generation takes place. To generate pseudo-labels, a confidence threshold is selected, and all predictions on the unlabeled dataset with probabilities greater than or equal to this threshold are used as pseudo-labels. In this study, several confidence threshold values ranging from 0.6 to 0.9 were evaluated.

In the next step, the pseudo-labels are incorporated into the labeled dataset in order to train a new instance of ResNet-18 from scratch. This process can be repeated iteratively to expand and refine the dataset. The iterative approach requires a termination condition, which can be defined as an integer indicating the number of iterations to perform. In this study, the termination condition was set to 14 iterations, which, combined with the initial supervised training, results in 15 total training iterations, each consisting of 100 epochs.

After each training iteration, the model's accuracy on the test dataset is evaluated. The model with the best accuracy is saved as checkpoint, and if a subsequent model performs worse, the saved model is loaded.

When the termination condition is met, the final model performance is evaluated on the test dataset.

3. Results

Throughout the study, various confidence thresholds were applied to each instance of the SSL approach, and their performance was subsequently evaluated on the test dataset.

3.1. Comparison of different confidence thresholds

In the experiment, confidence thresholds ranging from 0.6 to 0.9 were tested for SSL. For each threshold value, a new instance of the model was trained from scratch. Each instance yielded different accuracy results and changes in dataset size across iterations. Charts with results and the dataset size change for each instance are presented in Fig. 6.

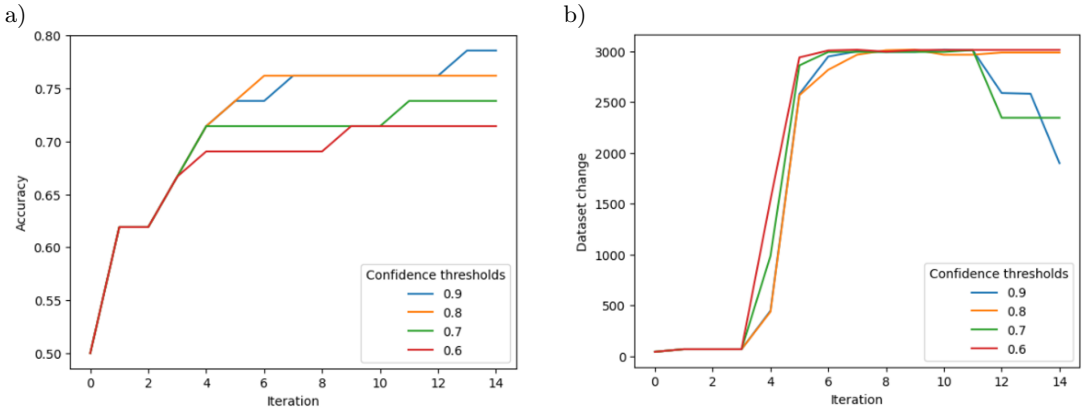

Fig. 6. Comparison of accuracy (a) and dataset size change (b) throughout each iteration of training for different confidence thresholds.

As shown in the charts, the overall accuracy on the test dataset ranges from 71.4 % to 78.6 %. The best performance was observed at a confidence threshold of 90 %, which includes only the most confident predictions as pseudo-labels. In comparison with other instance the results indicate that higher confidence thresholds are associated with improved model performance, likely due to the inclusion of more reliable pseudo-labels during training. Additionally, the dataset initially received very few pseudo-labels during the first three iterations of each

of the model instances. This was followed by a rapid increase to approximately 3 000 samples from the unlabeled dataset. In some cases, including the best-performing model instance, it discarded certain pseudo-labeled samples, likely to reduce noise and preserve label quality, thereby facilitating additional gains in performance.

3.2. Best model instance

Table 1 presents detailed information regarding the best-performing instance of the ResNet-18 model, obtained after completing all iterations of the SSL process with a confidence threshold set at 90 %. The table includes the confusion matrix with prediction information for each type of sediment and performance metrics such as: producer's accuracy (recall), user's accuracy (precision), omission error, commission error, overall accuracy, kappa, and $F1$-score.

Table 1. Confusion matrix for the best ResNet-18 instance evaluated on test dataset.

	Reference				**Sum**
	Boulders	Sand	Sandy gravel	Very fine sand	
Boulders	8	3	0	0	**11**
Sand	0	6	0	0	**6**
Sandy gravel	2	1	6	0	**9**
Very fine sand	1	2	0	13	**16**
Sum	**11**	**12**	**6**	**13**	**42**
Producer's accuracy	0.727	0.500	1.000	1.00	–
User's accuracy	0.727	1.000	0.667	0.813	–
Omission error	0.273	0.500	0.000	0.000	–
Commission error	0.273	0.000	0.333	0.187	–
Overall accuracy	0.786	–	–	–	–
Kappa	0.711	–	–	–	–
$F1$-score (weighted)	0.757	–	–	–	–

The confusion matrix reveals that the most challenging types of sediments to accurately predict is the category of boulders and sand classes. Specifically, boulders are frequently misclassified as sandy gravel and very fine sand, while the sand class itself is often confused with all of the other classes. This overlap makes sand the most difficult sediment type for the model to correctly identify. In contrast, sandy gravel and very fine sand were classified with relatively high accuracy, indicating better model performance on these classes. Other evaluation metrics further demonstrate the model's effectiveness, with an overall accuracy of 0.786 and a weighted $F1$-score of 0.757, reflecting solid predictive capability of the model.

4. Discussion

From this study, it can be concluded that selecting an appropriate confidence threshold during SSL has a noticeable impact on the final results. Each of the thresholds tested – from 60 % to 90 % led to different outcomes by the end of the SSL sessions. Although the differences in accuracy are not very large, ranging from 71.4 % to 78.6 %, the confidence threshold adjustment can still be significant, given that the same model architecture and hyperparameters were used consistently across all SSL experiments.

In terms of prediction details for the model using the optimal 90 % confidence threshold, several misclassification patterns were observed. Specifically, boulders were frequently misclassified as either sandy gravel or very fine sand. Additionally, the sand class posed a challenge, being commonly confused with all other sediment types. This class overlap made sand the most difficult sediment type for the model to accurately identify. In contrast, sandy gravel and very fine sand were classified with relatively high accuracy, indicating that the model performed better in these categories.

Misclassification of boulders as gravel may occur due to altered surface characteristics caused by red algae covering the boulders. Their presence increases the acoustic roughness and backscatter intensity of individual

boulders, which may be mistakenly interpreted as clusters of smaller rock fragments or finer sediments. Consequently, these features can be incorrectly classified as gravel accumulations rather than single boulders.

The misinterpretation of sandy seabed areas results from the specific nature of sand accumulations and their high susceptibility to seabed-shaping processes. In some locations, sand deposits form flat, even surfaces, while in others they form areas with megaripples or distinctly undulating relief. This variability in the morphology of sandy substrates can lead to divergent or inaccurate interpretations.

The analysis of sandy gravel and very fine sand yielded very good results. These areas are generally homogeneous, with limited morphological variability. Unlike sandy seabeds, which are often shaped by dynamic seabed processes. This consistency explains the relatively high accuracy achieved in the interpretation of gravelly-sand areas.

Future work could focus on some improvements to the performance of SSL for sediment classification. One promising direction is the incorporation of other SSL strategies, such as co-training (Chen *et al.*, 2024) or graph based SSL (Song *et al.*, 2021) which may also lead to more reliable results. Additionally exploring more sophisticated model architectures, such as Vision Transformers (Dosovitskiy *et al.*, 2021) or newer CNN architectures like EfficientNet (Tan, Le, 2019), may also offer performance gains. Finally, expanding the dataset with more diverse and balanced samples would help improve generalization of the models.

5. Conclusion

Sediment classification is an important topic of marine environmental research, which enables a better understanding of geological processes, habitat mapping, and the impact of human activities on underwater areas. The results obtained in this study confirm that the use of SSL methods is effective under conditions of limited availability of labeled data and a large volume of unlabeled data. By leveraging both annotated and unannotated samples, SSL approaches significantly reduce the dependence on labor-intensive and costly manual labeling, while maintaining competitive classification performance. A key conclusion that came out from this study is that the effectiveness of these methods depends on the confidence thresholds used to select pseudo-labeled samples, which directly influence the quality and reliability of the training process. The presented method increase the possibilities for practical application of deep learning in seafloor classification tasks and other domains where the acquisition of labeled samples is expensive, time-consuming, or logistically challenging. The findings highlight the potential of SSL as a cost-effective solution for advancing automated interpretation of seabed data.

Fundings

The work was supported by the National Science Center, Poland, project STREAMBAL No. 2021/41/B/ST10/01086.

Conflict of interest

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Authors' contributions

All authors contributed to the conceptualization of the study and participated in the analysis and interpretation of the results. Paweł Nadachowski was responsible for the study design. Maria Rucińska carried out data collection and pre-processing. The original draft of the manuscript was prepared by Paweł Nadachowski, Karolina Trzcińska, and Radosław Wróblewski. Revision of the manuscript was undertaken by Zbigniew Łubniewski, Jarosław Tęgowski, and Maria Rucińska. All authors agree to be accountable for all aspects of the work. All authors reviewed and approved the final manuscript.

References

1. Chen M., Du Y., Zhang Y., Qian S., Wang C. (2024), Semi-supervised learning with multi-head co-training, *arXiv*, http://arxiv.org/abs/2107.04795.
2. Dosovitskiy A. *et al.* (2021), An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv*, http://arxiv.org/abs/2010.11929.
3. Fonseca L., Calder B. (2005), Geocoder: An efficient backscatter map constructor, [in:] *Proceedings of the U.S. Hydro 2005 Conference*.
4. Hady M.F.A., Schwenker F. (2013), Semi-supervised learning, [in:] *Handbook on Neural Information Processing*, Bianchini M., Maggini M., Jain L. [Eds], **49**: 215–239, Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-36657-4_7.
5. He K., Zhang X., Ren S., Sun J. (2016), Deep residual learning for image recognition, [in:] *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
6. Kingma D.P., Ba J. (2017), Adam: A method for stochastic optimization, *arXiv*, http://arxiv.org/abs/1412.6980.
7. Krizhevsky A., Sutskever I., Hinton G.E. (2012), ImageNet classification with deep convolutional neural networks, [in:] *Proceedings of the 26th International Conference on Neural Information Processing Systems*.
8. Lee D.H. (2013), Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, [in:] *Workshop on Challenges in Representation Learning*.
9. Lemaître G., Nogueira F., Aridas C.K. (2017), Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning, *Journal of Machine Learning Research*, **18**(17): 1–5.
10. Parnum I.M., Gavrilov A.N. (2011), High-frequency multibeam echo-sounder measurements of seafloor backscatter in shallow water: Part 2 – Mosaic production, analysis and classification, *Underwater Technology*, **30**(1): 13–26, https://doi.org/10.3723/ut.30.013.
11. Paszke A. *et al.* (2019), PyTorch: An imperative style, high-performance deep learning library, [in:] *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.
12. Qin X., Luo X., Wu Z., Shang J. (2021), Optimizing the Sediment Classification of Small Side-Scan Sonar Images Based on Deep Learning, *IEEE Access*, **9**: 29416–29428, https://doi.org/10.1109/ACCESS.2021.3052206.
13. Schimel A.C.G. *et al.* (2018), Multibeam sonar backscatter data processing, *Marine Geophysical Research*, **39**: 121–137, https://doi.org/10.1007/s11001-018-9341-z.
14. Song Z., Yang X., Xu Z., King I. (2021), Graph-based semi-supervised learning: A comprehensive review, *arXiv*, http://arxiv.org/abs/2102.13303.
15. Tan M., Le Q. (2019), EfficientNet: Rethinking model scaling for convolutional neural networks, [in:] *Proceedings of the 36th International Conference on Machine Learning*, **97**: 6105–6114.
16. Zhao Y., Zhu K., Zhao T., Zheng L., Deng X. (2023), Small-sample seabed sediment classification based on deep learning, *Remote Sensing*, **15**(8): 2178, https://doi.org/10.3390/rs15082178.

Archives of Acoustics Vol. 51, No. 2, pp. 387–397 (2026), https://doi.org/10.24423/archacoust.2026.4337

OSA 2025

Application of ISO 12913 Standard to Assess Urban Soundscapes: A Case Study on Poznań

Jakub Dumanowski*, Anna Preis, Jan Felcyn

Department of Acoustics, Adam Mickiewicz University in Poznań
Poznań, Poland

*Corresponding Author: jakub.dumanowski@amu.edu.pl

Received September 17, 2025; revised January 20, 2026; accepted January 27, 2026; available online January 30, 2026; version of record April 16, 2026; published issue June 24, 2026.

ISO 12913 standards provide a unified framework for describing and assessing soundscapes, however, the lack of a Polish translation has so far limited their practical use. This paper presents the first application of a validated Polish version of the ISO/TS 12913-2 perceptual attributes, enabling full cross-language comparability of results. Whereas Polish research has traditionally focused on noise annoyance and broad judgements of acoustic comfort or discomfort, we outline the complete ISO-compliant assessment procedure, which combines: a soundwalk, questionnaires and audio-visual recording. The study was conducted at eight diverse urban locations in Poznań, Poland. Participants rated the soundscapes using eight attributes: *przyjemne*, *tętniące życiem*, *bogate w wydarzenia*, *chaotyczne*, *dokuczliwe*, *monotonne*, *ubogie w wydarzenia*, *spokojne*. Each rating set is mapped to a point in the 2D pleasantness-eventfulness space defined in ISO/TS 12913-3, facilitating visual comparison of locations and the identification of design needs. Results reveal pronounced perceptual differences between spatial typologies and demonstrate that the standardized approach provides richer, multidimensional information about the acoustic environment than conventional noise indicators. The proposed methodology establishes a reference framework for Polish soundscape studies and can support the creation of more people-friendly urban acoustic environments.

Keywords: soundscape, ISO 12913, Polish translation, soundwalk, perceptual attributes.

1. Introduction

International Organization for Standardization [ISO] 12913 standards provide a unified framework for describing and assessing soundscapes. Part 1 (ISO, 2014) defines the concept of a 'soundscape' and presents its conceptual model. Part 2 (ISO, 2018) specifies the requirements for data collection and reporting in soundscape studies, while part 3 (ISO, 2025) sets out methods for analyzing and interpreting those data. Method A in part 2 is a valuable source for acquiring quantitative data during soundwalks. The questionnaire permits a subjective evaluation of the perceived affective quality of the acoustic environment using eight attributes – pleasant, vibrant, eventful, chaotic, annoying, monotonous, and calm – on the five-point bipolar Likert scale. These attributes are embedded in the soundscape circumplex model (Axelsson *et al.*, 2010; ISO, 2025). In the ideal circumplex, adjacent attributes (i.e., pleasant–vibrant) are spaced 45° apart, whereas opposing ones (i.e., pleasant–annoying) are 180° apart (Fig. 1). From these eight attributes, the formulas in ISO/TS 12913-3 yield the indices pleasantness and eventfulness, which are displayed in a 2D eventfulness–pleasantness coordinate system (ISO, 2025; Mitchell *et al.*, 2022).

The Soundscape Attributes Translation Project (SATP) demonstrated that equal angular spacing between attributes is an idealized assumption and the angles depend strongly on the language in which the acoustic

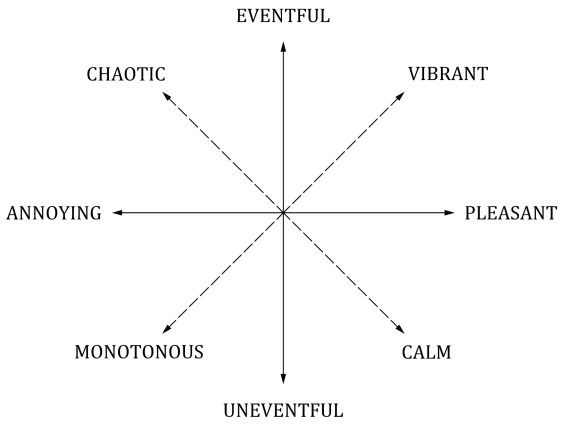

FIG. 1. Soundscape circumplex model adapted from Fig. A.1 of ISO/TS 12913-3:2025 (ISO, 2025).

environment is assessed (ALETTA *et al.*, 2024). The project developed a protocol for validating translations of the ISO (2018) soundscape attributes, consisting of a headphone-based listening experiment and a four-step validation method employing various statistical analyses. Another outcome was the update of ISO/TS 12913-3 (2025), which now includes correction angles for 13 languages that successfully passed validation, to be applied when calculating pleasantness and eventfulness. This update ensures cross-lingual comparability of soundscape assessments.

Pleasantness (P_{ISO}) and eventfulness (E_{ISO}) coordinates are calculated (ALETTA *et al.*, 2024; ISO, 2025) using:

$$P_{\mathrm{ISO}} = \frac{1}{\lambda_P} \sum_{i=1}^{8} \cos(\theta_i) \cdot \xi_i, \tag{1}$$

$$E_{\mathrm{ISO}} = \frac{1}{\lambda_E} \sum_{i=1}^{8} \sin(\theta_i) \cdot \xi_i, \tag{2}$$

where i indexes each circumplex scale, θ_i is the adjusted angle for the i-th soundscape attribute, and ξ_i is the value for that scale. The $1/\lambda$ provides a scaling factor to bring the range of P_{ISO}, E_{ISO} values to $[-1, +1]$:

$$\lambda_P = \frac{\rho}{2} \sum_{i=1}^{8} |\cos\theta_i|, \tag{3}$$

$$\lambda_E = \frac{\rho}{2} \sum_{i=1}^{8} |\sin\theta_i|, \tag{4}$$

where ρ is the range of the possible response values (i.e., $\rho = 5 - 1 = 4$ for the Likert scale, $\rho = 100$ for 0 to 100 scale responses).

2. Polish version of soundscape attributes

Until now, Polish psychoacoustic research has usually assessed soundscapes differently – by determining their annoyance, comfort or discomfort (PREIS *et al.*, 2015; SZYCHOWSKA *et al.*, 2018; FELCYN *et al.*, 2021). Although Polish studies using ISO 12913 exist (MŁYNARCZYK, WICIAK, 2024), the manner in which the individual attributes were translated in their questionnaires is unclear. The lack of a Polish version of ISO 12913 created the need for a validated Polish version of the soundscape attributes. Consequently, we contacted the SATP leadership to join the project as researchers from Adam Mickiewicz University in Poznań. Through our participation, we developed a validated Polish attribute set – *przyjemne, tętniące życiem, bogate w wydarzenia, chaotyczne,*

dokuczliwe, monotonne, ubogie w wydarzenia, spokojne (Dumanowski *et al.*, 2025) – and obtained the adjusted angles required to calculate pleasantness and eventfulness (Table 1). Thus, a methodology for proper soundscape assessment in the Polish language is now established.

Table 1. Polish translation of ISO (2018) soundscape attributes with obtained adjustment angles.

ISO (2018) soundscape attribute	ISO (2019) original angle [°]	Validated Polish translation	Obtained Polish adjustment angle [°]
Pleasant	0	Przyjemne	0
Vibrant	45	Tętniące życiem	69
Eventful	90	Bogate w wydarzenia	91
Chaotic	135	Chaotyczne	128
Annoying	180	Dokuczliwe	176
Monotonous	225	Monotonne	266
Uneventful	270	Ubogie w wydarzenia	274
Calm	315	Spokojne	339

It should be noted that the correction angles affect only the transformation of raw attribute assessments into the pleasantness–eventfulness circumplex and not the soundwalk procedure or the perceptual judgments. Pleasantness and eventfulness are calculated using language-specific correction angles to ensure cross-language and cross-cultural compatibility, while variations in angle values affect only the numerical positioning within the 2D space.

The subsequent sections of this paper present the procedure and results of the first pilot soundwalk employing the validated Polish attributes and the calculation of pleasantness and eventfulness using the Polish correction angles.

3. Methods

3.1. Soundwalk route

On 13 May 2025 a soundwalk was carried out in the center of Poznań, Poland, under dry, calm weather conditions (wind speed below 3 m/s, temperature 18.5 °C, relative humidity 38 %). The route comprised eight evaluation points (see Fig. 2) and ran from the Kaponiera Roundabout to the Chrobry Bridge. The first stop, P1, was the large, traffic-intensive Kaponiera Roundabout (*Rondo Kaponiera*); P2 was Mickiewicz Square (*Plac Mickiewicza*) on St. Martin Street; P3 led into Mickiewicz Park (*Park Mickiewicza*), a green space with a fountain on Fredro Street. P4, Freedom Square (*Plac Wolności*), is another central plaza with a fountain, while P5, Old Market Square (*Stary Rynek*), represents historical center of the city. From there the walk continued to P6,

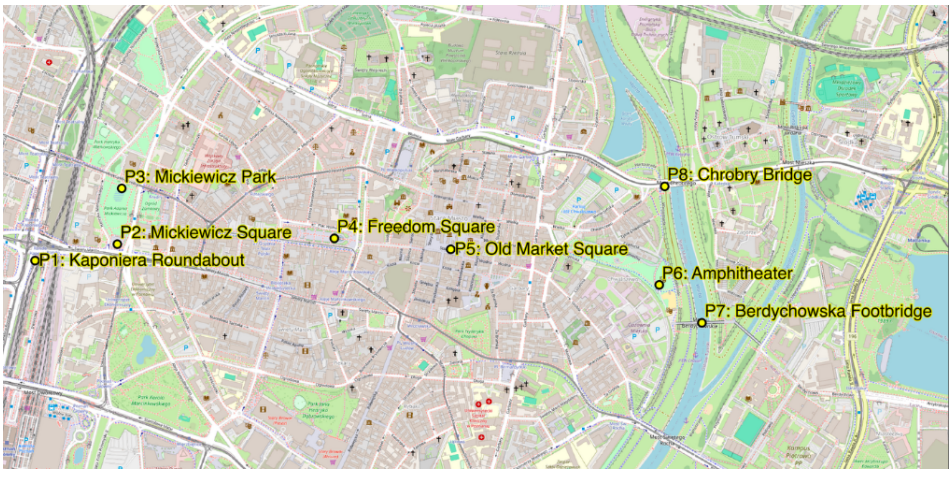

Fig. 2. Soundwalk points in Poznań on map derived from OpenStreetMap.

the Amphitheater (*Amfiteatr*) in the riverside park located next to the cultural-recreational KontenerART area, proceeded across P7, the Berdychowska Footbridge (*Kładka Berdychowska*) over the Warta River, and ended at P8, the Chrobry Bridge (*Most Chrobrego*), which spans the Warta River and links the heavily trafficked Estkowski Street and Wyszyński Street.

3.2. Participants

Thirteen participants (5 females, 8 males; age range 22 to 73 years; M_{age} = 29, SD_{age} = 14.4) took part in the soundwalk. The group consisted of acoustics students along with three lecturers from Adam Mickiewicz University.

3.3. Procedure

At each point the participants evaluated the soundscape using the Polish-language soundscape questionnaire translated from ISO (2018). The survey was hosted online: participants scanned a QR code that redirected them to a pre-prepared questionnaire created using the FreeOnlineSurveys. Within the form (Fig. 3) they identified audible sound sources, rated the eight soundscape attributes and could enter free comments. All ratings were given on interactive sliders ranging from 0 to 100. The structure of our questionnaire was inspired by the survey used in the article by Mitchell *et al.* (2020).

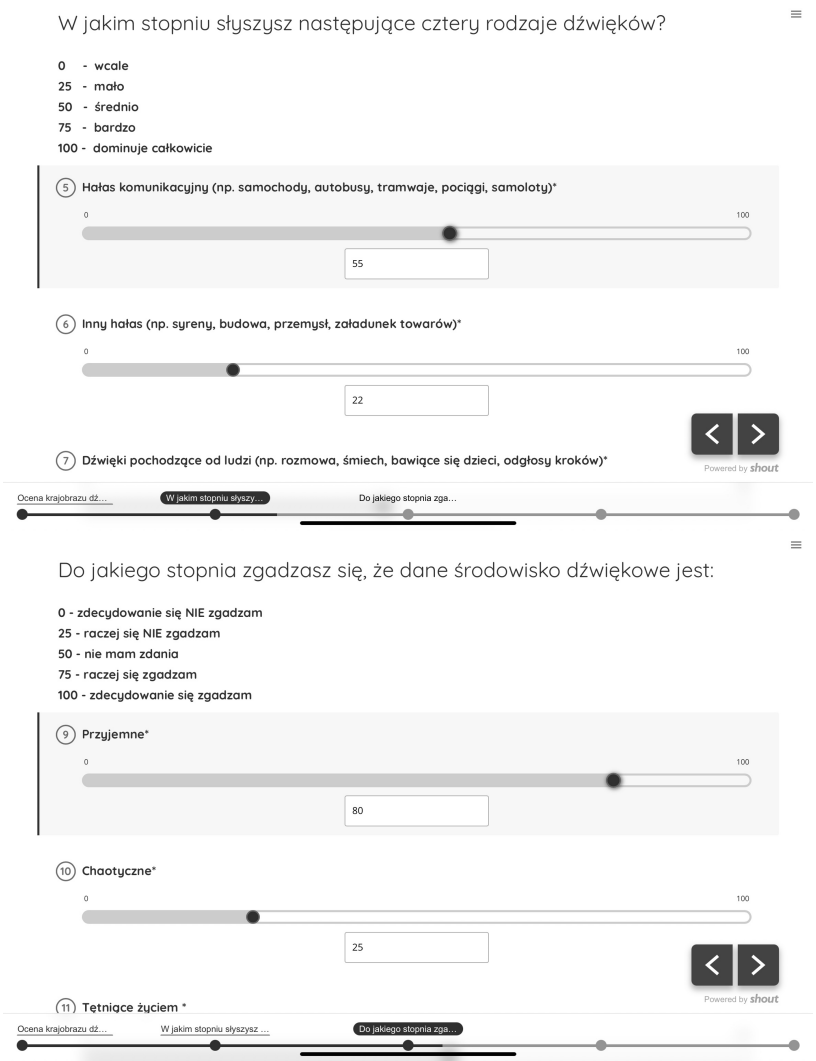

Fig. 3. Screenshots of the graphical user interface for evaluating soundscape using FreeOnlineSurveys.

For sound-source identification the question read: 'To what extent do you presently hear the following four types of sounds? (0 – not at all, 100 – dominates completely).' The four categories presented were: traffic noise (e.g., cars, buses, trams, trains, airplanes), other noise (e.g., sirens, construction work, industrial activity, loading of goods), human sounds (e.g., conversation, laughter, children playing, footsteps), and natural sounds (e.g., birdsong, flowing water, wind in vegetation).

For the attribute assessment (pleasant, vibrant, eventful, chaotic, annoying, monotonous, uneventful, calm) it read: 'To what extent do you agree or disagree that the present surrounding sound environment is...? (0 – strongly disagree, 100 – strongly agree).' The soundscape evaluation at each location lasted approximately 5 min.

While the questionnaire was being completed, binaural audio, ambisonic audio and 360° video were recorded simultaneously (see Fig. 4 for the recording setup). A calibrated recording device (HEAD acoustics, SQuadriga II – Mobile recording and playback system) with binaural microphones (HEAD acoustics, BHS II – Binaural Headset for Aurally Accurate Recording and Playback) was used, enabling subsequent extraction of equivalent sound levels and psychoacoustic parameters from the recorded samples. The visual environment was captured using a 360° camera (GoPro MAX 360 Action Camera), while ambisonic audio was recorded using the first-order ambisonic microphone (RØDE NT-SF1 Ambionic Microphone) with a multichannel audio recorder (Zoom F6 Field Recorder).

Fig. 4. Recording setup – binaural microphones, ambisonic microphone, and 360° video camera.

4. Results

4.1. Participants' subjective soundscape evaluations

Based on the ratings of the eight soundscape attributes, the indices pleasantness and eventfulness were computed using Eq. (1) to Eq. (4). Figure 5 plots every single assessment (all participants at all eight points) to illustrate the spread across the 2D eventfulness–pleasantness space. Figure 6 shows the individual eventfulness–pleasantness ratings for the eight Poznań locations made by the 13 soundwalk participants, together with the median value for each site. Kernel-density contours representing the 10th, 25th, 50th, and 75th percentiles are superimposed to visualize the concentration of responses. Figure 7 presents the mean perceived prominence of the four predefined sound-source categories at each location; error bars indicate the 95 % confidence intervals.

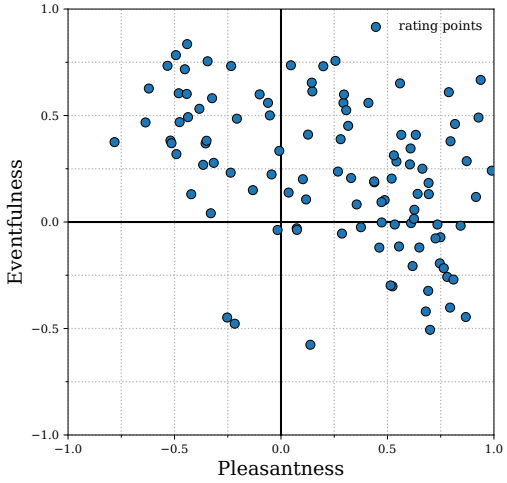

FIG. 5. All participants' ratings at all eight locations mapped onto eventfulness–pleasantness coordinate system.

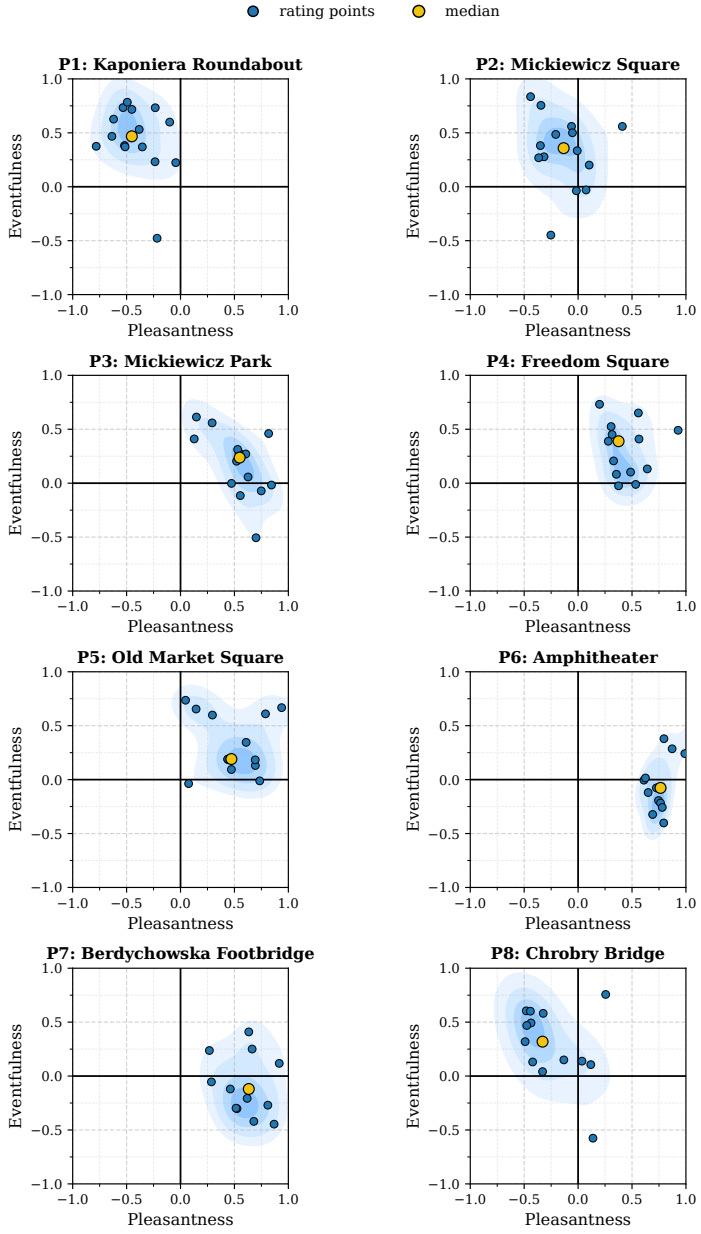

FIG. 6. Eventfulness–pleasantness ratings for each of the eight locations in Poznań.

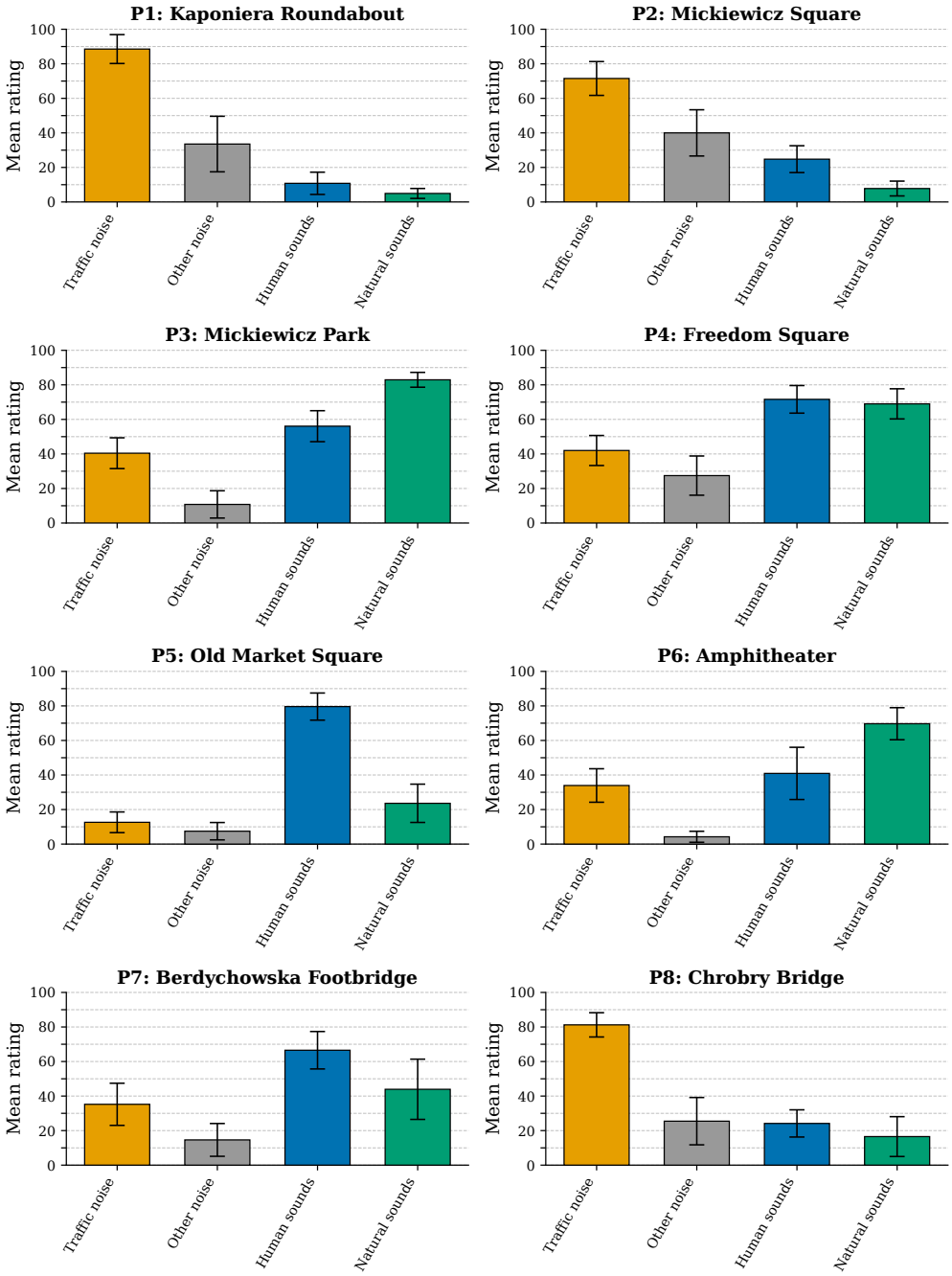

Fig. 7. Mean value of perceived prominence of the four predefined sound-source categories at each location.

4.2. Objective parameters calculated from binaural recordings

In accordance with the requirements of ISO (2019), objective acoustical parameters – equivalent sound level (L_{Aeq}), loudness, N5, N95, sharpness, fluctuation strength, roughness and tonality – were extracted from approximately five-minute binaural recordings at each of the eight measurement points using standard-compliant sound analysis software (HEAD acoustics, ArtemiS SUITE Software (Datasheet)). Because binaural recordings provide separate left- and right-ear channels, the channels were processed individually. In line with the standard, the higher of the two values was retained for every descriptor. The values of all calculated objective parameters are listed in Table 2. A visual representation of this data is shown in Fig. 8.

TABLE 2. Objective parameters of eight evaluated locations, calculated from binaural recordings.

Location	P1: Kaponiera Roundabout	P2: Mickiewicz Square	P3: Mickiewicz Park	P4: Freedom Square	P5: Old Market Square	P6: Amphitheater	P7: Berdychowska Footbridge	P8: Chrobry Bridge
L_{Aeq} [dB]	85.9	66.9	68.3	59.1	61.8	53.8	55.8	72.3
Loudness [sone]	53.8	24.4	24.7	15.2	16.4	10.0	11.3	29.6
N5 [sone]	45.1	25.9	26.9	18.4	16.2	10.7	13.8	35.7
N95 [sone]	12.5	13.9	21.2	11.4	10.2	5.2	5.2	11.6
Sharpness [acum]	2.45	2.02	3.94	2.60	2.15	1.69	1.80	2.44
Fluctuation strength [vacil]	0.08	0.05	0.01	0.01	0.02	0.01	0.02	0.01
Roughness [asper]	0.04	0.05	0.03	0.02	0.03	0.02	0.03	0.04
Tonality [tuHMS]	0.21	0.26	0.14	0.16	0.18	0.17	0.15	0.16

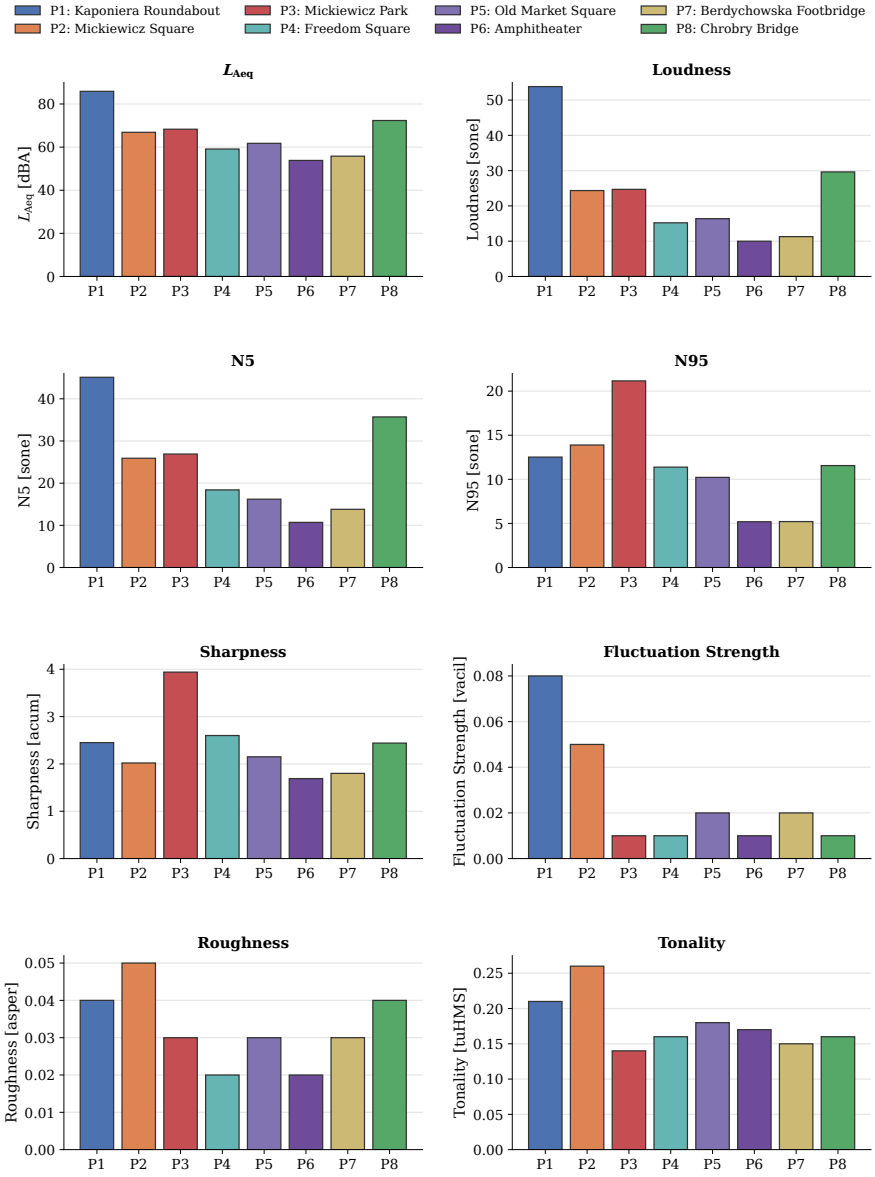

FIG. 8. Objective parameters across eight evaluated locations.

P1: Kaponiera Roundabout shows the highest values of L_{Aeq}, loudness, N5, and fluctuation strength, as well as the highest median eventfulness, while recording the lowest median Pleasantness. P3: Mickiewicz Park has the greatest background loudness (N95) and the highest sharpness value, and its soundscape contains the largest share of natural sounds. At P2: Mickiewicz Square, traffic noise dominates, yet among all eight sites this square also contains the highest proportion of 'other' noises; it exhibits the greatest roughness and tonality. At P5: Old Market Square, human sounds represent the largest share of the soundscape. The most favorable soundscape is found in P6: Amphitheater, where L_{Aeq}, loudness, N5, N95, sharpness, and roughness reach their lowest values, and pleasantness is the highest of all locations. The lowest median Eventfulness is observed on P7: Berdychowska Footbridge.

5. Discussion

The individual-level data reveal a considerable spread, an expected consequence of the subjective nature of the ratings. One way to tighten the dispersion could be to inform participants in advance how to interpret each soundscape attribute; however, such instruction could introduce response bias. Although ISO/TS 12913-2 recommends a minimum of 20 respondents, the present study was conceived as a pilot intended to test the in-situ applicability of the Polish attribute set.

All judgments were made on a continuous 0–100 slider rather than on the five-point Likert scale suggested by ISO (2018). The finer 101-point resolution offers greater numerical precision when computing pleasantness and eventfulness. While this would be impractical with paper forms, the online survey interface made the slider implementation straightforward. A future experiment could explicitly compare the 0–100 slider with the five-step Likert format.

A few participants scored eventfulness markedly differently from the majority, possibly because the Polish terms 'bogate w wydarzenia' and 'ubogie w wydarzenia' were misunderstood, or because momentary lapses of attention led to reversed ratings.

As expected, soundscapes dominated by traffic noise received lower pleasantness scores than those characterized by human voices or natural sounds, confirming earlier findings (Nilsson, Berglund, 2006; Nilsson *et al.*, 2007; Axelsson *et al.*, 2010). According to Schafer's (1993) typology, the sites studied can be classified as hi-fi environments (P3: Mickiewicz Park, P4: Freedom Square, P5: Old Market Square, P6: Amphitheater, P7: Berdychowska Footbridge) and lo-fi environments (P1: Kaponiera Roundabout, P2: Mickiewicz Square, P8: Chrobry Bridge). In general, high L_{Aeq} and high loudness are associated with low pleasantness, whereas low values of these measures coincide with high pleasantness. This relationship is clear in very quiet and very loud contexts, where N5, loudness, and L_{Aeq} are good predictors of the pleasantness. In the mid-range of sound levels, the pattern weakens and exceptions emerge. For instance, P3: Mickiewicz Park was rated more pleasant than P2: Mickiewicz Square even though it showed higher L_{Aeq}, loudness, N5, N95, and sharpness, probably due to the fountain's masking effect and the presence of human voices and natural sounds. In contrast to level-related metrics, parameters describing temporal and tonal sound characteristics, such as fluctuation strength, sharpness, roughness, and tonality, were not significantly associated with either pleasantness or eventfulness.

These results indicate that sound level alone is insufficient to predict soundscape quality. They support the view that 'informational properties of soundscapes (i.e., categories of sounds) are better predictors of perceived soundscape quality than acoustic measures such as L_{Aeq}' (Axelsson *et al.*, 2010; Nilsson, 2007).

6. Conclusions

The study presented an evaluation of eight locations in Poznań during a pilot soundwalk conducted in accordance with ISO (2018), using the validated Polish version of the soundscape attributes. The proposed methodology establishes a reference framework for Polish soundscape studies and can guide the design of more people-friendly urban acoustic environments. Future work should recruit a larger and more diverse participant pool (beyond

individuals linked to acoustics) and include sites that are monotonous. Follow-up studies might also apply the soundscape assessment protocol in laboratory settings to complement the in-situ findings.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Authors' contributions

Jakub Dumanowski conceptualized the study, wrote the original draft, prepared the surveys and the sound walk, developed the methodology, curated the data, performed the analysis, and created the visualizations. Anna Preis developed the methodology, prepared and supervised the sound walk, and conceptualized the study. Jan Felcyn performed the analysis, curated the data, and contributed to data interpretation. All authors reviewed and approved the final manuscript.

Acknowledgments

The authors would like to thank Eryk Kozłowski for his assistance with the organization and the audiovisual recordings, and the first-year Acoustics students of Adam Mickiewicz University in Poznań for participating in the soundwalk.

References

1. Aletta F. *et al.* (2024), Soundscape descriptors in eighteen languages: Translation and validation through listening experiments, *Applied Acoustics*, **224**: 110109, https://doi.org/10.1016/j.apacoust.2024.110109.
2. Axelsson Ö., Nilsson M.E., Berglund B. (2010), A principal components model of soundscape perception, *The Journal of the Acoustical Society of America*, **128**(5): 2836–2846, https://doi.org/10.1121/1.3493436.
3. Dumanowski J., Felcyn J., Preis A. (2025), Polish version of soundscape attributes: Translation process and preliminary validation results, [in:] *Proceedings of the 11th Convention of the European Acoustics Association Forum Acusticum / EuroNoise 2025*, pp. 4823–4826, https://doi.org/10.61782/fa.2025.0684.
4. Felcyn J., Preis A., Praszkowski M., Wrzosek M. (2021), Assessment of audio-visual environmental stimuli. Complementarity of comfort and discomfort scales, *Archives of Acoustics*, **46**(2): 279–288, https://doi.org/10.24425/aoa.2021.136582.
5. International Organization for Standardization (2014), *Acoustics – Soundscape. Part 1: Definition and conceptual framework* (ISO Standard No. ISO 12913-1:2014), https://www.iso.org/standard/52161.html.
6. International Organization for Standardization (2018), *Acoustics – Soundscape. Part 2: Data collection and reporting requirements* (ISO Standard No. ISO/TS 12913-2:2018), https://www.iso.org/standard/75267.html.
7. International Organization for Standardization (2019), *Acoustics – Soundscape. Part 3: Data analysis* (ISO Standard No. ISO/TS 12913-3:2019), https://www.iso.org/standard/69864.html.
8. International Organization for Standardization (2025), *Acoustics – Soundscape. Part 3: Data analysis* (ISO Standard No. ISO/TS 12913-3:2025), https://www.iso.org/standard/86955.html.
9. Mitchell A. *et al.* (2020), The Soundscape Indices (SSID) Protocol: A method for urban soundscape surveys–questionnaires with acoustical and contextual information, *Applied Sciences*, **10**(7): 2397, https://doi.org/10.3390/app10072397.

10. Mitchell A., Aletta F., Kang J. (2022), How to analyse and represent quantitative soundscape data, *JASA Express Letters*, **2**(3): 037201, https://doi.org/10.1121/10.0009794.

11. Młynarczyk D., Wiciak J. (2024), Virtual reality technology in analysis of the Sarek National Park soundscape in Sweden, *Archives of Acoustics*, **49**(3): 319–329, https://doi.org/10.24425/aoa.2024.148802.

12. Nilsson M.E., Berglund B. (2006), Soundscape quality in suburban green areas and city parks, *Acta Acustica united with Acustica*, **92**(6): 903–911.

13. Nilsson M.E. (2007), A-weighted sound pressure level as an indicator of perceived loudness and annoyance of road-traffic sound, *Journal of Sound and Vibration*, **302**(1–2): 197–207, https://doi.org/10.1016/j.jsv.2006.11.010.

14. Nilsson M.E., Botteldooren D., De Coensel B. (2007), Acoustic indicators of soundscape quality and noise annoyance in outdoor urban areas, [in:] *Proceedings of the 19th International Congress on Acoustics (ICA 2007)*, https://doi.org/1854/10363.

15. Preis A., Kociński J., Hafke-Dys H., Wrzosek M. (2015), Audio-visual interactions in environment assessment, *Science of the Total Environment*, **523**: 191–200, https://doi.org/10.1016/j.scitotenv.2015.03.128.

16. Schafer R.M. (1993), *Our Sonic Environment and the Soundscape: The Tuning of the World*, Destiny Books, Rochester.

17. Szychowska M., Hafke-Dys H., Preis A., Kociński J., Kleka P. (2018), The influence of audio-visual interactions on the annoyance ratings for wind turbines, *Applied Acoustics*, **129**: 190–203, https://doi.org/10.1016/j.apacoust.2017.08.003.

Archives of Acoustics Vol. 51, No. 2, pp. 399–413 (2026), https://doi.org/10.24423/archacoust.2026.4327

OSA 2025

The Impact of Generated and Expressive Modulation of the Synthetic Instrument Sound Parameters on the Impression of Naturalness

Marek Pluta

Department of Mechanics and Vibroacoustics, AGH University of Krakow
Kraków, Poland

e-mail: pluta@agh.edu.pl

Received September 8, 2025; revised March 4, 2026; accepted March 8, 2026;
available online March 11, 2026; version of record June 3, 2026; published issue June 24, 2026.

Despite their different spectral structures, the sound of early instruments from the electrophone group was often considered to be deceptively similar to the sound of wind or bowed string instruments. However, the wavetable synthesizer playing a short, looped sample of a natural instrument is easily distinguishable from the actual instrument. This results from the presence of specific modulatory structures in the sound of some instruments related to expression, which can be a strong clue regarding the identification of the instrument. The control of early electrophones, such as the theremin or Martenot waves, gave the performer expressive capabilities comparable to bowed instruments. Contemporary synthesizers are returning to similar solutions. The aim of this work is to study the impact of various types of modulation on the perceived naturalness of violin sound. Modulation through an automatic low frequency oscillator is compared to expressive modulation by a human using a controller. Two advanced controllers are studied to determine whether simultaneous modulation of more than one parameter brings benefits. A set of sound samples was prepared which included violin recordings and synthesized signals, where different waveforms were combined with various modulation sources and modulated parameters. The effect was assessed by a group of expert listeners. The results indicate that expressive, multi-parameter modulation with advanced controllers brings benefits for waveforms with realistic spectra, close to that of a violin. In less realistic waveforms this kind of modulation may be perceived as less natural than a simple one, obtained through an oscillator.

Keywords: sound synthesis, signal modulation, expressive performance, expressive controller.

1. Introduction

Due to their versatility, sound synthesizers are often considered as a replacement for acoustic instruments. Such scenarios require a great deal of effort to reproduce the sound features of the original instrument. The bare minimum is to produce a sound that can be recognised as a chosen instrument, while the ultimate goal is to convince a listener that the sound is not synthetic, but original. A set of features (McAdams, Bruno, 2012) that need to be reproduced depends on the instrument. For some instruments, it is enough to concentrate on spectral features, while others require a proper reproduction of particular temporal structures, or combination of both (Iverson, Krumhansl, 1993).

The violin, along with the remaining members of the bowed string group, is a particularly difficult case, where synthetic counterparts are still easily distinguished from the original. The difficulty is caused by the continuous control that a performer has over multiple sound parameters. However, this specific way of controlling sound parameters is a means to produce a unique, and highly expressive performance. A key to achieve a convincing performance with synthesizers is therefore not only the ability of a synthesizer to allow control over multiple sound parameters, but also the source of a control data.

The sound of the early instruments belonging to the electrophone group was often considered to be deceptively similar to the sound of wind instruments or bowed strings, even though its spectral features differed from the original. On the other hand, the wavetable synthesizer playing a short, looped sample recorded from a natural instrument, could be easily distinguished from the actual instrument due to the absence of some modulatory structures specific for the instrument and related to expression. These structures may be a stronger clue regarding the identification of the instrument than the sound spectrum. The control of early electrophones, such as gestures used in theremin, or a moving keyboard, a ring, and a touch-sensitive lozenge in Martenot waves, gave the performer expressive capabilities comparable to bowed and wind instruments. Contemporary synthesizers are introducing control solutions based on similar assumptions, allowing to combine multiple parameters in one gesture. This gives an opportunity to study a combination of artificial sound source with natural, expressive modulation.

The synthesis of a violin sound is a known issue, and various attempts have been made to improve it. In a recent study (Liu, 2024), a violin sound has been synthesized using the wavetable synthesis method with the vibrato effect simulated using pitch modulation. The modulation rate and depth varied according to simple few-segment envelopes. The envelope parameters were set according to the performance statistics described in (Schoonderwaldt, Friberg, 2001). Another study (Kim *et al.*, 2025) proposes to model the natural pitch contour using a two-stage diffusion-based synthesis framework. The first stage is responsible for the estimation of the fundamental frequency contour that controls the Musical Instrument Digital Interface (MIDI) pitch bend. The second stage generates Mel spectrogram that applies these expressive details. The attempt is similar to a prior study (Wu *et al.*, 2022) that uses differentiable digital signal processing (DDSP) to generate expressive deviations of the synthesis parameters.

Another approach to the problem, based strongly on the physics of instruments, has been proposed for the case of spectral synthesis techniques (Pérez Carrillo, 2009). The extensive study presents solutions to issues such as measuring violin performance using recorded sounds, as well as devices recording bow motion and force, establishing the relationship between performance data and sound timbre, and designing a generative model of timbre. The findings have been applied to design an advanced sound synthesizer. However, the applied model of vibrato is relatively simple. It uses a sine modulation of pitch, with depth and rate controlled by a fade-in and fade-out envelope with random deviations based on measurements.

The above presented studies are based either on numerical modelling of physical objects or on some form of machine learning. The comprehensive review of both approaches (Hawley *et al.*, 2020) discusses their capabilities to achieve realistic sound features. The conclusion is that in order to produce convincing sounds of instruments it is not necessary to conduct a detailed physical simulation, which requires a further validation based on measurements with live musicians and real instruments. Instead, very good results can be obtained when either a human expert or a deep neural network selects and tunes a set of salient factors based on knowledge or a very large set of audio recordings.

All of the approaches discussed above present a view of the problem focused mainly on the side of the sound source. They analyse models of instruments and refinements that can be applied to various methods of sound synthesis. A view from a different perspective, centred around human perception of sound, can provide valuable insight into the problem (Fritz *et al.*, 2025). The study compares an actual human performance using a real instrument to a bowing machine playing a real violin, and a hybrid sound synthesis that takes control data from a bowing force recorded with a human violinist. The goal of the comparison is to capture properties perceived in the actual instruments. The main part of the study discusses the methodologies based on listening tests that are applied to evaluate perceived sound qualities. The hybrid method yielded the best results, allowing one to test unlimited variants of instrument properties with the same natural excitation, and showing the best correlations with acoustic measurements. This shows that in the case of studying violin, both controlled playing conditions and natural excitation of the instrument are equally important. It can be seen that player interaction can influence the perception of produced sound even more than the properties of the instrument.

Current research trends appear to focus on one particular use case, where a synthesizer serves as a means to automatically produce an artificial recording of an instrument on the basis of the musical score alone. Such scenario

requires an implementation of a simulated expression. There is, however, another use case, where a synthesizer is used in a live performance, using some form of controller, such as a keyboard. Here, all the expression can, and should, come from the performer. There are, however, two limitations: synthesis parameters available to be assigned with the expression data, and limitations of the controller. A simple keyboard does not allow one to control the parameters in a continuous manner. An addition of aftertouch capability, that is, a continuous sensitivity to a force applied to the already depressed key solves the problem. However, if more than one parameter needs to be controlled, which is the case of a violin, the aftertouch is insufficient. Thus, more advanced controllers are required.

Various experiments have been carried out with designs adapted for specific instruments, such as accordion (Gurevich, von Muehlen, 2001) or even for human voice (Donati, Chousidis, 2022). Some attempts have been made with game controllers adapted for controlling sound synthesis parameters, but finally expressive extension[1] for the original MIDI specification[2] allowed one to design general purpose multiple degrees of freedom controllers for synthesizers (Robertson, 2011). These controllers attempt to give the musician a possibility to freely manipulate multiple parameters, possibly with a single gesture. Typical solutions include touch-sensitive keys that react not only to a change in pressure, but also to a change of finger placement on the key surface, which allows a single finger to control three parameters. Other solutions include keys with additional rotational axes.

When combined with a sufficiently complex synthesizer, advanced, multiple degree of freedom controllers have a potential to significantly improve features of a synthesized sound by adding a natural expressive performance. So far, this possibility has not undergone a thorough study, and remains open.

The aim of this work is to study the impact of various types of modulation on the perceived naturalness of violin sound, and to compare modulation through an automatic low-frequency oscillator to expressive modulation by a human using a controller. Two advanced controllers are studied to determine whether simultaneous modulation of more than one parameter brings benefits. A synthesizer with architecture open for controller-related modifications has been designed as a base for the current study and its future continuation. Using the synthesizer, a set of sound samples was prepared which included violin recordings and synthesized signals, where different waveforms were combined with various modulation sources and modulated parameters. The effect was assessed by a group of expert listeners. The results indicate that expressive, multi-parameter modulation with advanced controllers brings benefits for waveforms with realistic spectra, close to that of a violin. However, in less realistic waveforms, this kind of modulation may be perceived as less natural than a simple one, such as that obtained through an oscillator.

2. Experiment

The aim of the experiment was to present a group of expert listeners with a set of sound samples. The samples contained a middle section of a single violin note, a^1 (fundamental frequency, $f_0 = 440\,\mathrm{Hz}$). They represented combinations of various synthesized sounds with several variants of modulation. Some modulations were generated using low frequency oscillator (LFO), some were recorded by a violinist using an expressive controller. The listeners were asked to assess how natural each sample was. The analysis of listeners' assessments should give information regarding the impact of particular sound features on the impression of naturalness of a violin sound.

Figure 1 presents the procedure and elements of the experiment. The main element is a program that serves three purposes. It is a sound synthesizer, and plays a sound in response to data from the controller. The controller, operated by a violinist, produces a stream of modulation data, which is recorded by the program. With data from the controller recorded, the program can use it to modulate selected synthesis parameters and record the generated audio signal. The supervisor chooses a desired base signal and adds LFO or recorded modulation to the selected parameters. The resulting audio signals are recorded to be used in the listening test.

[1] MIDI 1.0 Detailed Specification, 1996.
[2] MIDI Polyphonic Expression, 2018.

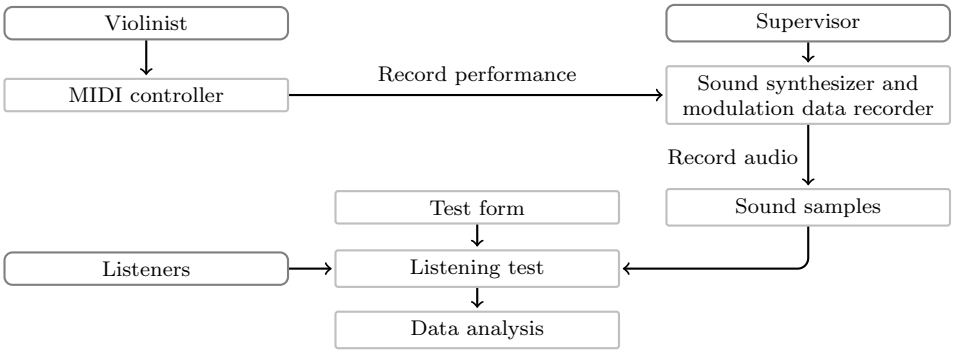

Fig. 1. Overall diagram of the experiment.

2.1. Synthesis and modulation

The program is implemented in the Max/MSP visual programming language, which is capable of handling data from MIDI Polyphonic Expression (MPE)-compatible controllers. It consists of three modules. The first is responsible for real-time performance and reacts to data from the controller by synthesizing a sound and adjusting its parameters accordingly, providing the performer with auditory feedback.

The second module catches and stores a stream of MPE data sent by the controller. This data is the recording of an expressive performance by the violinist. Three separate streams can be recorded simultaneously: pitch bend data, amplitude data, and aftertouch data. The pitch bend data controls the instantaneous deviation of the fundamental frequency, which is a main component of the violin vibrato. The amplitude stream is independent of the single-number MIDI velocity parameter, sent when a key is pressed. It can control the amplitude envelope or add amplitude modulation which, to some extent, usually occurs in parallel with frequency modulation. The aftertouch data are interpreted as a timbre parameter, which controls the cut-off frequency of the low-pass filter applied to the signal. Again, it may be used to add an envelope or oscillatory modulation.

The last program module uses the same synthesizer as the first real-time module to reproduce a sound with selected modulation. The source of modulation can be an internal LFO, or a stored performance stream. The selected signal is combined with the selected modulation source and applied to selected parameters: pitch, amplitude, or cut-off frequency. The synthesized sound is played and stored in an audio file.

The synthesizer can produce three types of constant signal: sawtooth, filtered sawtooth, and a single looped period of a real violin sound, recorded in an anechoic chamber using a close microphone (Fig. 2). The first two may be considered as case of subtractive synthesis. The last one is based on the wavetable principle and reproduces the real violin spectrum, but is devoid of any internal evolution. The filtered sawtooth uses three resonant low-pass filters to roughly shape lower regions of the spectrum, similarly to the violin spectrum. Thus, the three signal types are graded from purely synthetic (sawtooth), to a violin-like (wavetable). The selected waveform passes through a controllable low-pass resonant filter and a controllable amplifier. The frequency deviation of the waveform, the filter cut-off frequency, and the signal amplitude can be modulated either by a common internal LFO, or separate data streams recorded from the controller and implemented as envelope generators, as shown in Fig. 3.

Violin recordings, carried out in an anechoic chamber, were analysed and used as a reference for setting ranges for modulation parameters. They were compared and verified against values found in (Schoonderwaldt, Friberg, 2001). The modulation parameters are presented in Table 1.

Table 1. Ranges of modulation parameters.

Modulation source	Modulation rate	Pitch-bend	Amplitude	Filter cut-off [Hz]
LFO	6 Hz	±25 cent	±1 dB	3000 to 6000
Controller	Variable	±25 cent	100 %	3000 to 6000

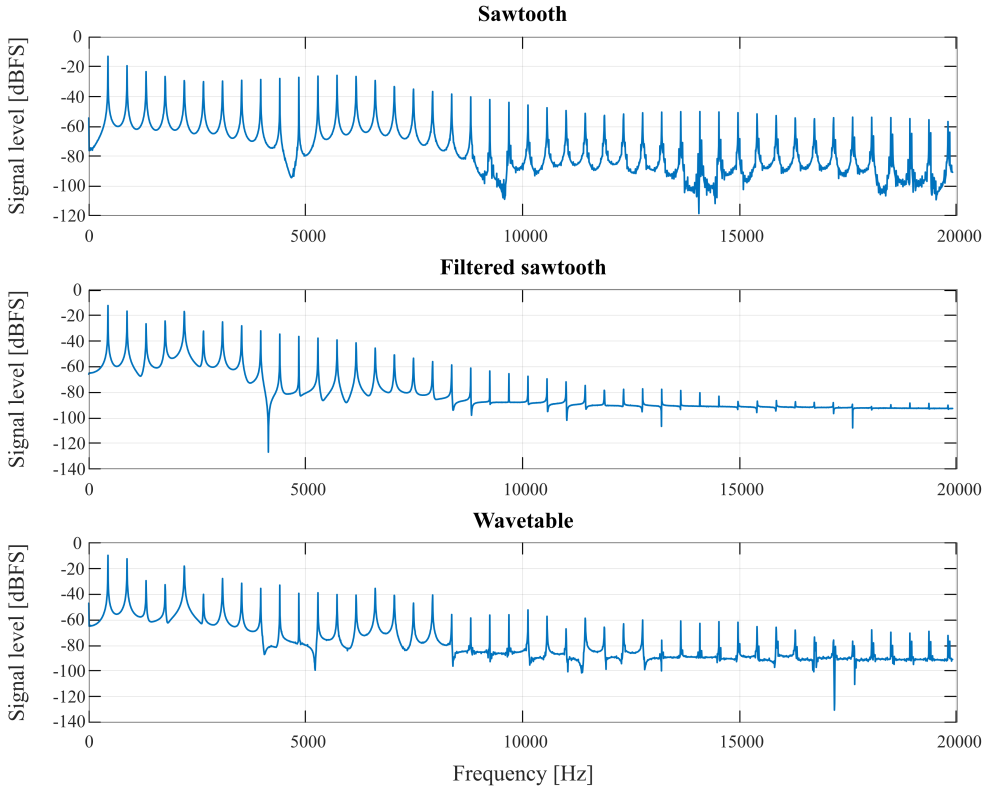

FIG. 2. Spectra of the synthesized signals. Base waveforms are filtered with a default setting of aftertouch-controlled low-pass resonant filter.

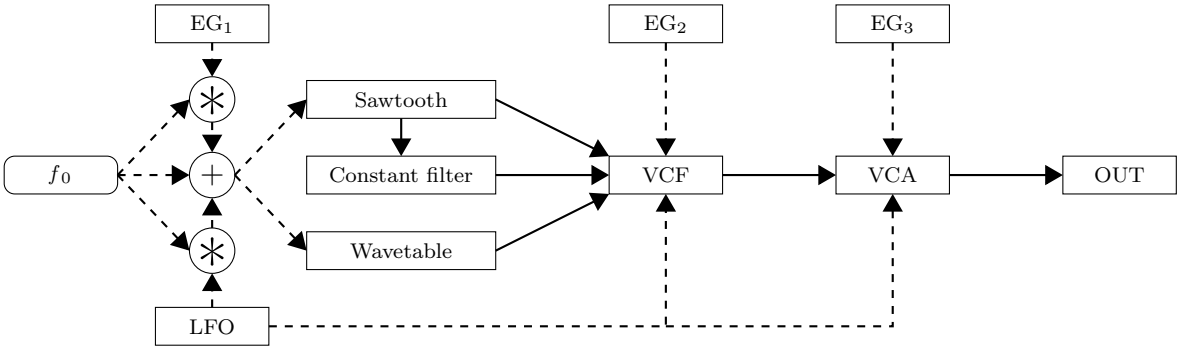

FIG. 3. Diagram of the synthesizer (VCF (voltage controlled filter) – adjustable low-pass filter, VCA (voltage controlled amplifier) – adjustable amplifier). Pitch, filter cut-off, and amplitude are modulated either by internal LFO, or separate envelope generators (EG_1–EG_3). One of three available waveforms is used at a time.

2.2. CONTROLLERS

Performance data streams were recorded using two controllers: Expressive E Osmose, and Joue Play. The former is a novel keyboard controller. The latter is an universal controller with interchangeable templates. Both allow controlling three parameters with a single finger, although each has its own limitations.

Expressive E Osmose (Fig. 4) is a keyboard with an additional axis: the keys can be pressed down and moved sideways. Pressing is divided into two regions: in the first region, the key moves lightly and in the second region there is a perceptible key resistance. The boundary between both regions is easily perceptible and each region controls its own parameter. In a default mode, which has been used in research, the light region controls the amplitude, the resistant region controls the aftertouch, and the sideways movement controls the pitch bend. A clear limitation is the inability to simultaneously change the amplitude and aftertouch, which belong to different regions of the same key travel.

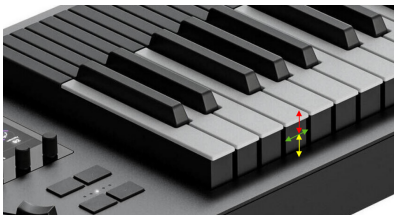

Fig. 4. Controller Expressive E Osmose (green arrow shows pitch bend control, red – amplitude control, yellow – aftertouch).

Joue Play (Fig. 5) was used with a template consisting of 17 identical key-straps. Each strap works as an XY touch-sensitive pad. Moving a finger sideways controls the pitch bend, moving it upwards or downwards controls the aftertouch, and the pressure of the finger controls the amplitude. All three parameters can be controlled simultaneously. However, the actual resolution and sensitivity varies on the surface, which limits the combinations of parameter values achievable.

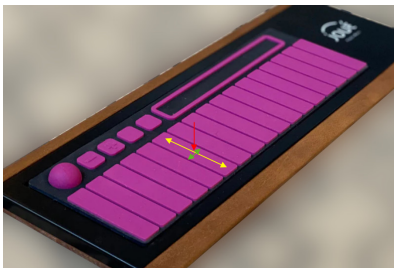

Fig. 5. Controller Joue Play (green arrow shows pitch bend control, red – amplitude control, yellow – aftertouch).

2.3. Recorded performance

The person using both controllers was a violinist experienced in bowed strings and electronic keyboard instruments. The synthesizer was set to reproduce the violin wavetable. The task of the musician was to perform a single, long note that would be as close as possible to the sound of the violin, with regard to expression. The violinist used an auditory feedback to refine the sound, until a satisfactory performance had been recorded. The recorded data streams are shown in Fig. 6. The amplitude and aftertouch are represented by unsigned 7-bit integer values. Pitch bend has a theoretical 14-bit resolution and is represented with floating point values.

The limitations of both controllers are clearly visible. Osmose is used in the aftertouch range; therefore, simultaneous amplitude control is impossible, and its value stays constant, apart from the moments of pressing and releasing a key. Joue seriously limits the practical aftertouch range due to large finger movement (along the entire length of a strap), and its amplitude stream shows some discontinuities.

2.4. Sound samples

Using both internal LFO and recorded performance data for modulation of selected parameters, a total of 67 different sound samples has been created. The samples differed in the choice of the base waveform, the modulation source, and the set of modulated parameters. Selection included:

1. Recording of the original instrument from an anechoic chamber.
2. 3 synthesized samples (different base waveforms) without modulation.
3. 63 synthesized samples with modulation – all combinations of the following variants:
   - 3 waveforms: violin wavetable (Vn), filtered sawtooth (Sf), and non-filtered sawtooth (Sn),
   - 3 modulation sources: Joue (J), Osmose (O), internal LFO (L),
   - 7 combinations of modulated parameters: amplitude (A), filter cut-off frequency (T – timbre), fundamental frequency (P – pitch).

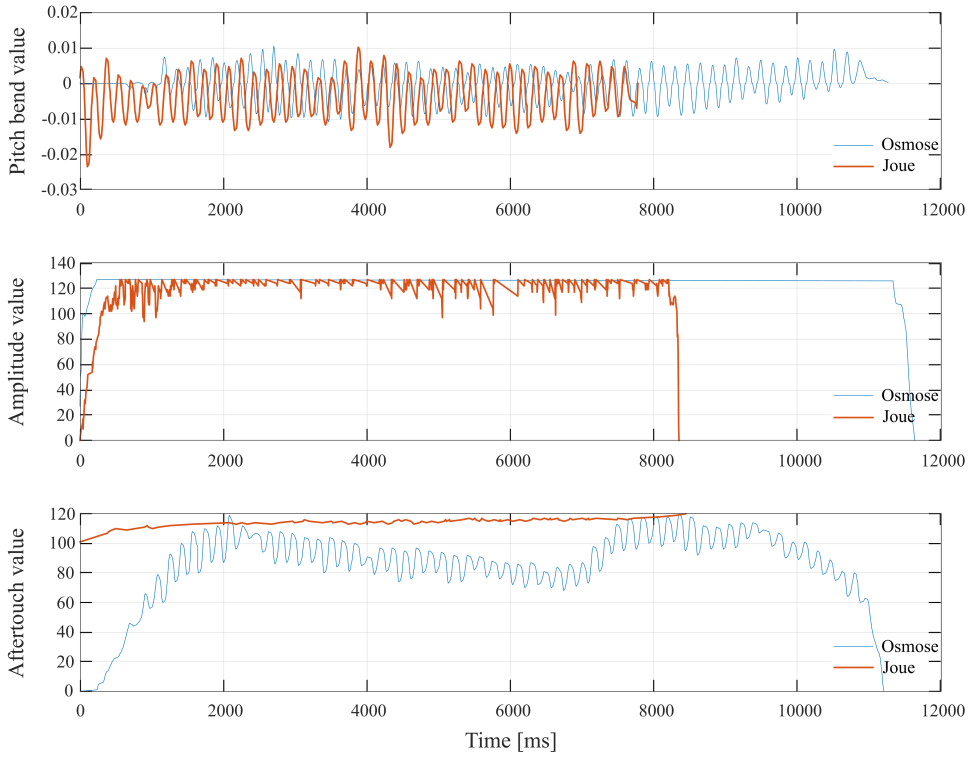

FIG. 6. Raw data recorded from controllers. Pitch bend controls signal fundamental frequency, amplitude controls signal level, and aftertouch controls filter cut-off frequency.

Despite the fact that modulation of the amplitude and filter cut-off frequency affected the signal level, sound samples were kept at their recorded levels and were not level-matched in loudness to reflect the original impact of expression on the signal. Excluding the original violin recording, the maximum peak level difference among sound samples reached 1.24 dB and the maximum RMS level difference reached 4.54 dB. Including the violin recording, the maximum peak level difference reached 6 dB and the maximum RMS level difference reached 10.74 dB.

Only 3 second fragments from the middle section of the recordings were selected for the test to exclude the influence of the attack phase, which was not relevant to the study. 50 ms fade in and fade out have been applied to all samples. Examples of three sound samples[3] are presented in Fig. 7, Fig. 8, and Fig. 9.

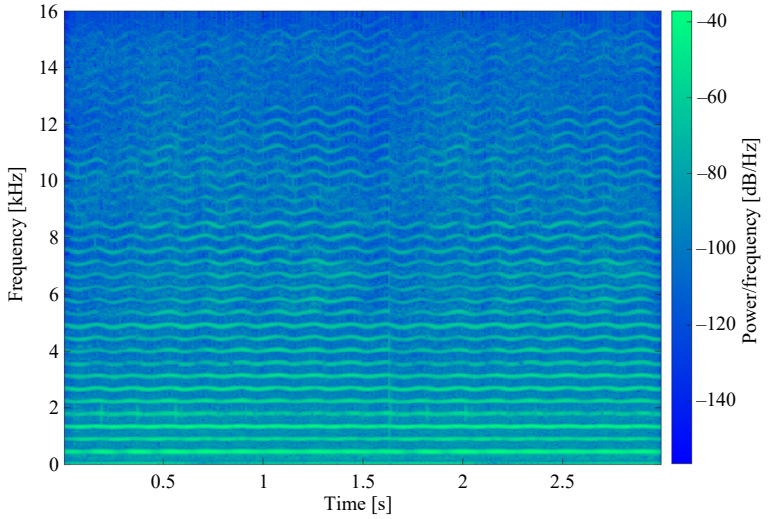

FIG. 7. Spectrogram of the violin recording.

[3]All sound samples are available for download.

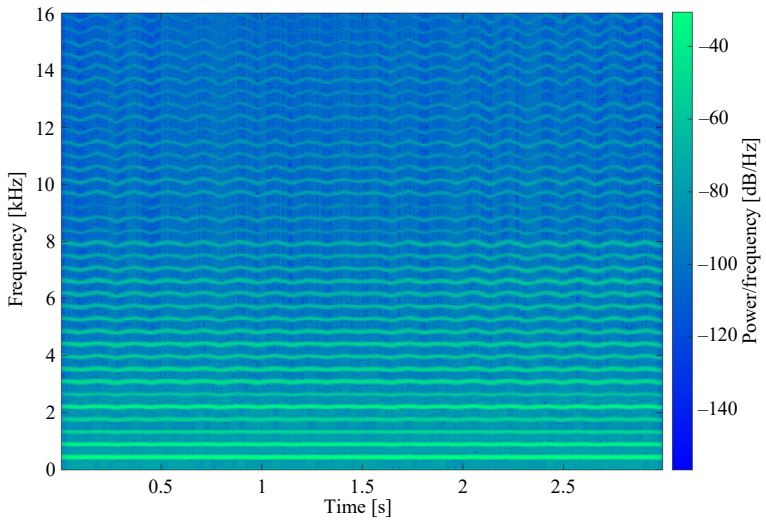

FIG. 8. Spectrogram of the violin wavetable; amplitude, cut-off, and pitch modulated by Osmose.

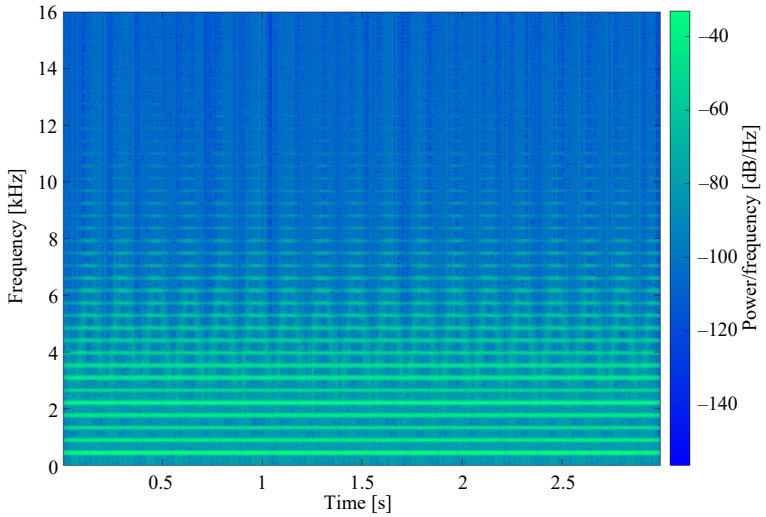

FIG. 9. Spectrogram of the filtered sawtooth; amplitude and cut-off modulated by LFO.

2.5. Listening test

The effect of modulation was evaluated by a group of expert listeners. They rated the naturalness of each sound sample using a 5 grade scale: excellent (5), good (4), fair (3), poor (2), and bad (1). The listeners were not informed about the details of sound processing. They were only asked to evaluate the samples as the sustain phase of a violin sound. It was possible to listen to the samples multiple times.

The responses were collected using an online form. The form included four questions regarding the nature of the listeners' experience: 1) experience playing the violin, 2) experience playing other instruments or singing, 3) experience in audio signal processing, and 4) experience in sound recording or music production. Additionally, five sound samples occurred twice in the test in order to evaluate coherence of the listener's responses. All samples were arranged in random order.

A total of 16 listeners participated in the test. Every listener had at least some experience (up to 4 years) in at least 2 categories or high experience in at least one category. 13 listeners had high experience in at least 2 categories, and 6 listeners had high experience in 3 categories. This translated well into coherence of responses. If a response to a repeated question differed by not more than 1, such a response was considered coherent. All listeners were coherent in at least 4 of 5 repeated questions, and 14 were coherent in all repeated questions. Therefore, all responses in the test were considered valid.

3. Results

All results have been presented in a form of histograms of the responses of the listeners. Histograms are divided into cases that represent combinations of various waveforms, modulation sources, and modulated parameters. Symbols used in charts are explained in Subsec. 2.4. All the scores presented in histograms are summarised in tables as means and medians.

For the baseline, the results for samples without any modulation for three types of waveforms are presented in Fig. 10 and in Table 2. As expected, more resemblance to the spectrum of the violin yields better results, although even the results for the violin wavetable are not better than fair.

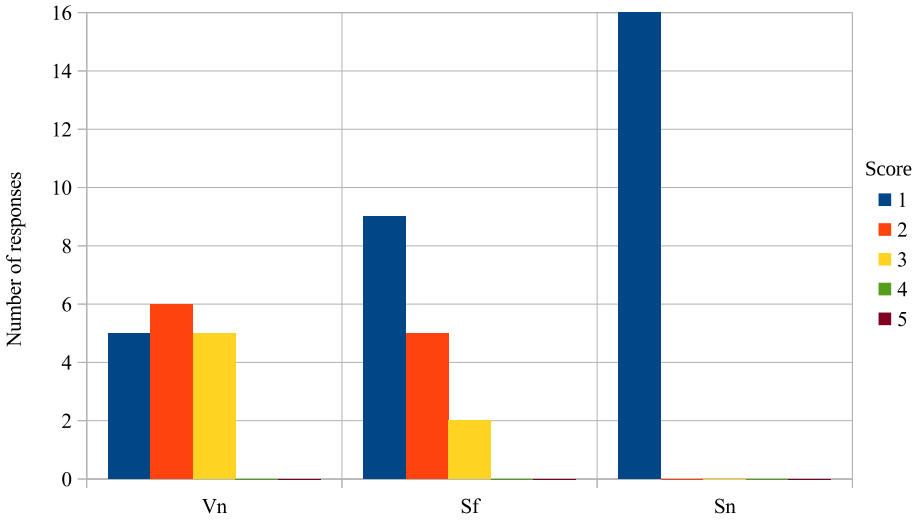

Fig. 10. Responses for all waveform types without modulation.

Table 2. Summary of ratings for all waveform types without modulation.

Sample	Mean score	Median
Vn	2.00 ±0.79	2.00
Sf	1.56 ±0.70	1.00
Sn	1.00 ±0.00	1.00

The second baseline is the violin recording, which in Fig. 11 and in Table 3 is compared to the most complete modulation of all three parameters (ATP), with cases representing three different modulation sources: two con-

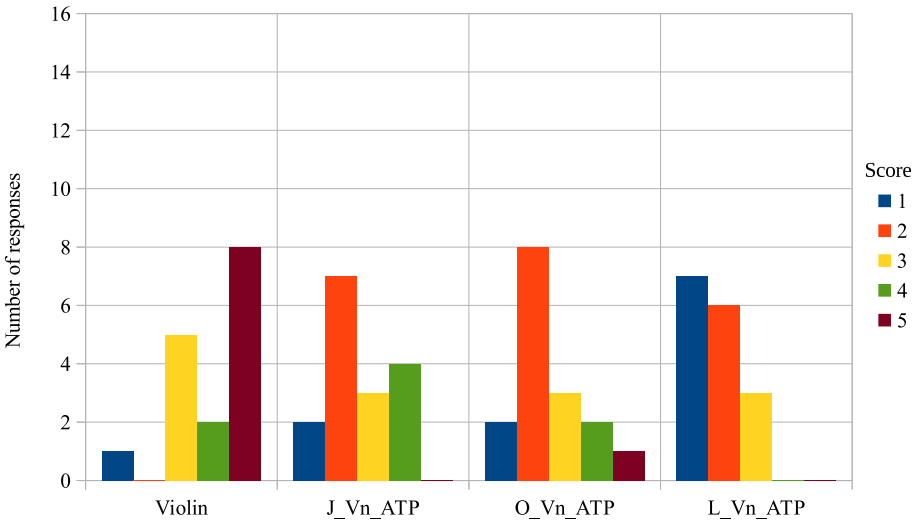

Fig. 11. Responses for the violin recording and for cases of complete, three parameter modulation (ATP).

Table 3. Summary of ratings for the violin recording and for cases of complete, three parameter modulation (ATP).

Sample	Mean score	Median
Violin	4.00 ±1.17	4.50
J_Vn_ATP	2.56 ±1.00	2.00
O_Vn_ATP	2.50 ±1.06	2.00
L_Vn_ATP	1.75 ±0.75	2.00

trollers and LFO. For the violin, half of the responses give the maximal score (5), which cannot be matched by any synthetic source. At the same time, it can be seen that with full modulation controllers yield better results than LFO, which is the least natural. The results of both controllers are similar.

Figure 12 to Fig. 14 and Table 4 present a complete set of the results obtained. Unsurprisingly, almost all responses for non-filtered sawtooth are bad. This kind of waveform cannot convince listeners even with all three parameters (ATP) modulated with expressive controllers. Both controllers with modulation of three parameters are the best cases here, as well as Osmose with pitch vibrato supplemented with modulation of the second parameter. However, most of the responses are bad. When comparing filtered sawtooth (Sf) to the wavetable (Vn), a larger spread of results is seen in case of the wavetable.

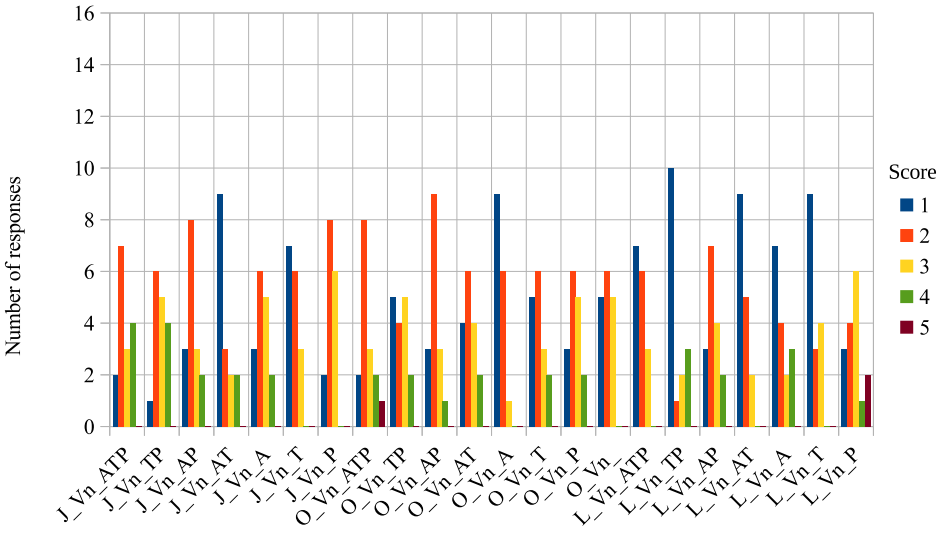

Fig. 12. Responses for all cases of wavetable waveform (Vn).

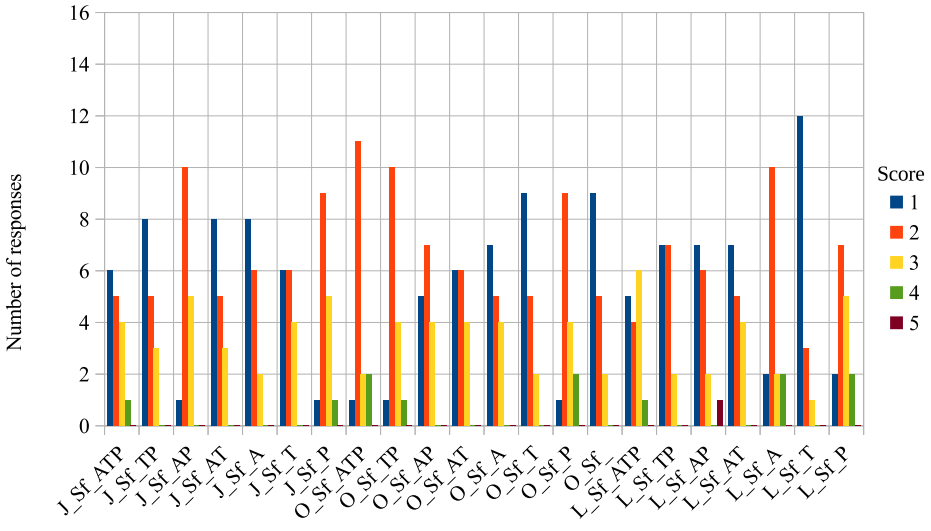

Fig. 13. Responses for all cases of filtered sawtooth waveform (Sf).

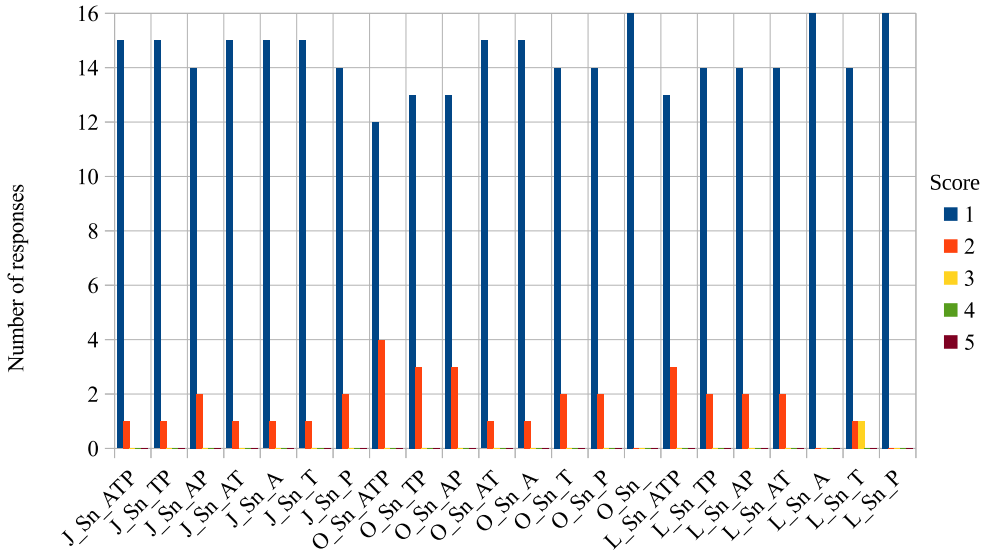

FIG. 14. Responses for all cases of non-filtered sawtooth waveform (Sn).

TABLE 4. Summary of ratings for all samples.

Sample	Mean score	Median	Sample	Mean score	Median	Sample	Mean score	Median
J_Vn_ATP	2.56 ±1.00	2.00	J_Sf_ATP	2.00 ±0.94	2.00	J_Sn_ATP	1.06 ±0.24	1.00
J_Vn_TP	2.75 ±0.90	3.00	J_Sf_TP	1.69 ±0.77	1.50	J_Sn_TP	1.06 ±0.24	1.00
J_Vn_AP	2.25 ±0.90	2.00	J_Sf_AP	2.25 ±0.56	2.00	J_Sn_AP	1.13 ±0.33	1.00
J_Vn_AT	1.81 ±1.07	1.00	J_Sf_AT	1.69 ±0.77	1.50	J_Sn_AT	1.06 ±0.24	1.00
J_Vn_A	2.38 ±0.93	2.00	J_Sf_A	1.63 ±0.70	1.50	J_Sn_A	1.06 ±0.24	1.00
J_Vn_T	1.75 ±0.75	2.00	J_Sf_T	1.88 ±0.78	2.00	J_Sn_T	1.06 ±0.24	1.00
J_Vn_P	2.25 ±0.66	2.00	J_Sf_P	2.38 ±0.70	2.00	J_Sn_P	1.13 ±0.33	1.00
O_Vn_ATP	2.50 ±1.06	2.00	O_Sf_ATP	2.31 ±0.77	2.00	O_Sn_ATP	1.25 ±0.43	1.00
O_Vn_TP	2.25 ±1.03	2.00	O_Sf_TP	2.31 ±0.68	2.00	O_Sn_TP	1.19 ±0.39	1.00
O_Vn_AP	2.13 ±0.78	2.00	O_Sf_AP	1.94 ±0.75	2.00	O_Sn_AP	1.19 ±0.39	1.00
O_Vn_AT	2.25 ±0.97	2.00	O_Sf_AT	1.88 ±0.78	2.00	O_Sn_AT	1.06 ±0.24	1.00
O_Vn_A	1.50 ±0.61	1.00	O_Sf_A	1.81 ±0.81	2.00	O_Sn_A	1.06 ±0.24	1.00
O_Vn_T	2.13 ±0.99	2.00	O_Sf_T	1.56 ±0.70	1.00	O_Sn_T	1.13 ±0.33	1.00
O_Vn_P	2.38 ±0.93	2.00	O_Sf_P	2.44 ±0.79	2.00	O_Sn_P	1.13 ±0.33	1.00
O_Vn	2.00 ±0.79	2.00	O_Sf	1.56 ±0.70	1.00	O_Sn	1.00 ±0.00	1.00
L_Vn_ATP	1.75 ±0.75	2.00	L_Sf_ATP	2.19 ±0.95	2.00	L_Sn_ATP	1.19 ±0.39	1.00
L_Vn_TP	1.88 ±1.22	1.00	L_Sf_TP	1.69 ±0.68	2.00	L_Sn_TP	1.13 ±0.33	1.00
L_Vn_AP	2.31 ±0.92	2.00	L_Sf_AP	1.88 ±1.05	2.00	L_Sn_AP	1.13 ±0.33	1.00
L_Vn_AT	1.56 ±0.70	1.00	L_Sf_AT	1.81 ±0.81	2.00	L_Sn_AT	1.13 ±0.33	1.00
L_Vn_A	2.06 ±1.14	2.00	L_Sf_A	2.25 ±0.83	2.00	L_Sn_A	1.00 ±0.00	1.00
L_Vn_T	1.69 ±0.85	1.00	L_Sf_T	1.31 ±0.58	1.00	L_Sn_T	1.19 ±0.53	1.00
L_Vn_P	2.69 ±1.21	3.00	L_Sf_P	2.44 ±0.86	2.00	L_Sn_P	1.00 ±0.00	1.00

In Fig. 15 and Fig. 16 and in Table 5 a selection of cases is presented to illustrate the impact of multi-parameter modulation, separately for the filtered sawtooth and for the wavetable. For both waveforms, the case of no modulation is the least natural. When modulation in the filtered sawtooth is applied only to pitch, which is the simplest form of vibrato, LFO and both controllers score very similar results. Strangely, for three-parameter modulation results for LFO and one controller (Joue) turn worse, mostly due to increased number of bad (1) scores. The results of the second controller (Osmose) with modulated three parameters do not change much in comparison to the modulation of one parameter. Clearly, multi-parameter modulation does not help with a waveform that is not very similar to the real instrument. The situation changes with the wavetable waveform. Here, transition from one parameter to three parameter modulation improves the results for both

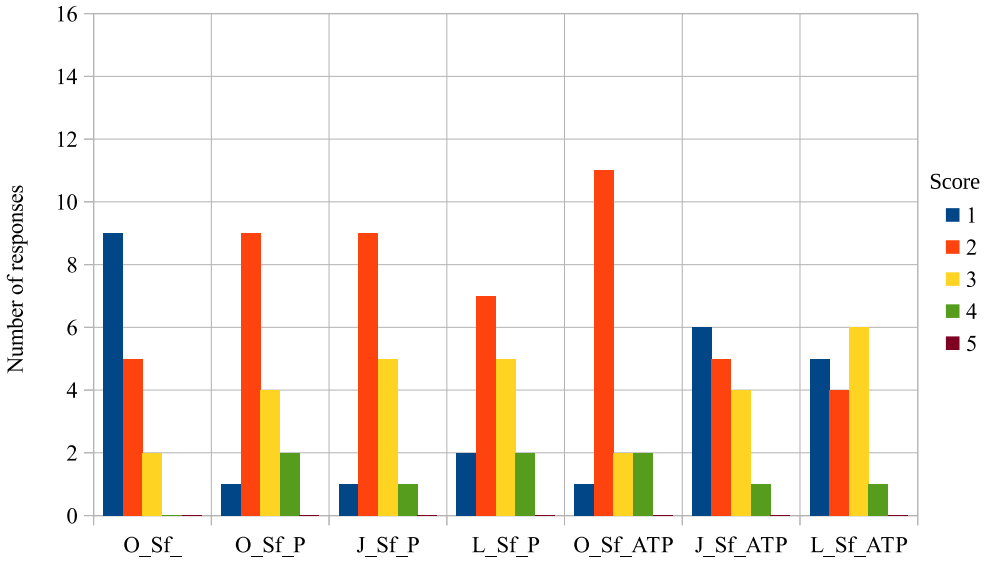

FIG. 15. Responses for cases of zero, one, and three parameters modulated in the filtered sawtooth waveform.

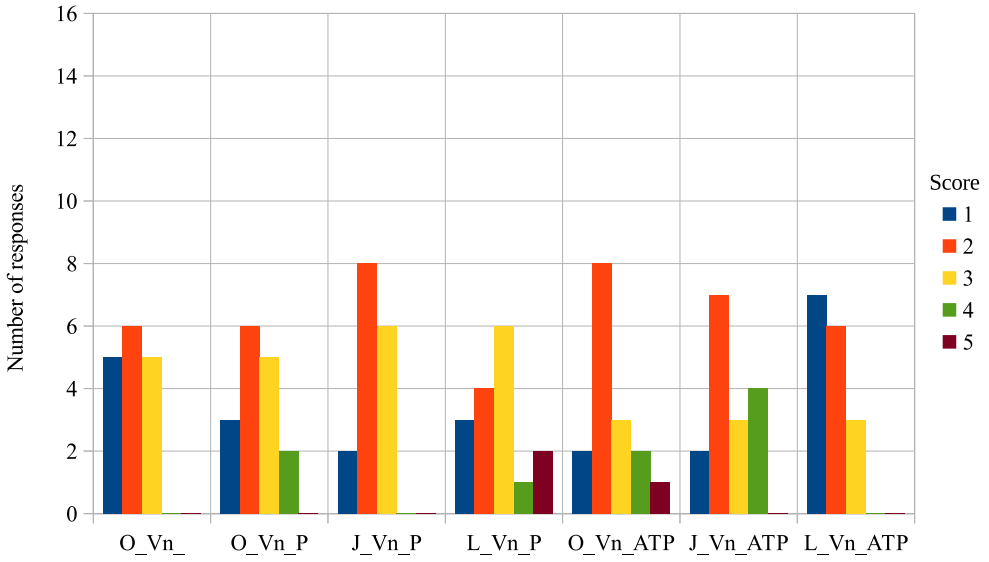

FIG. 16. Responses for cases of zero, one, and three parameters modulated in the wavetable waveform.

TABLE 5. Summary of ratings for cases of zero, one and three parameters modulated in the filtered sawtooth and in the wavetable waveform.

Sample	Mean score	Median	Sample	Mean score	Median
O_Sf	1.56 ±0.70	1.00	O_Vn	2.00 ±0.79	2.00
O_Sf_P	2.44 ±0.79	2.00	O_Vn_P	2.38 ±0.93	2.00
J_Sf_P	2.38 ±0.70	2.00	J_Vn_P	2.25 ±0.66	2.00
L_Sf_P	2.44 ±0.86	2.00	L_Vn_P	2.69 ±1.21	3.00
O_Sf_ATP	2.31 ±0.77	2.00	O_Vn_ATP	2.50 ±1.06	2.00
J_Sf_ATP	2.00 ±0.94	2.00	J_Vn_ATP	2.56 ±1.00	2.00
L_Sf_ATP	2.19 ±0.95	2.00	L_Vn_ATP	1.75 ±0.75	2.00

controllers, and worsens for the LFO. However, LFO yielded better results than controllers for single-parameter (pitch) modulation. With three parameters modulated Osmose is considered the most natural. The obvious conclusion is that any modulation makes the violin sound more natural. However, for expressive controllers in order to show their benefits, the spectrum of the signal needs to be similar to that of the original instrument. Only such cases are able to convince some respondents to reward the synthetic signal with the best score (5).

Figure 17 shows how the improvement of the waveform impacts the score when only the pitch is modulated. Here, gains for the case of LFO modulation are the largest. One of the controllers (Joue) sees a decrease in the score when the filtered sawtooth is changed to wavetable. Everything changes when three parameters are modulated (Fig. 18 and Table 6). In this case, the controllers produce the best scores for the wavetable, but the LFO score decreases.

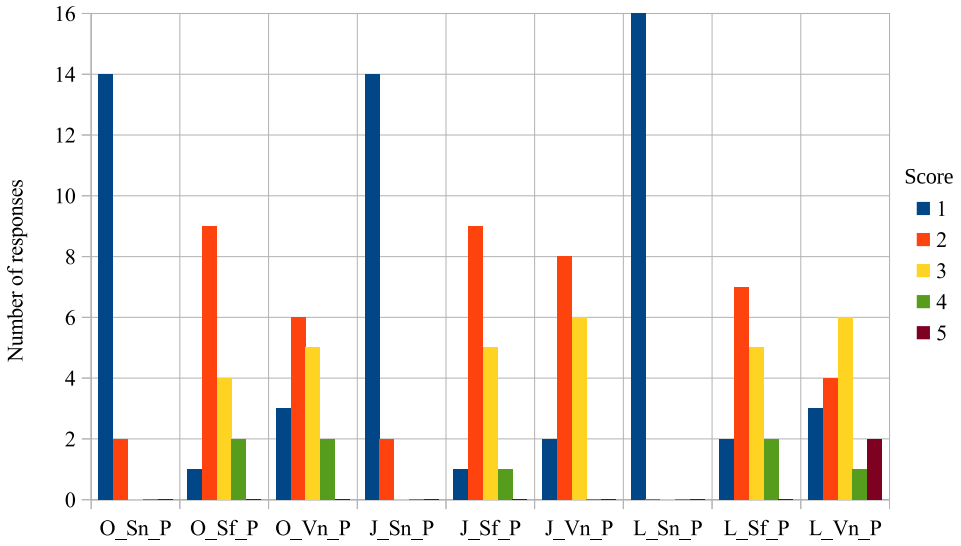

FIG. 17. Responses for various waveforms with pitch modulation only.

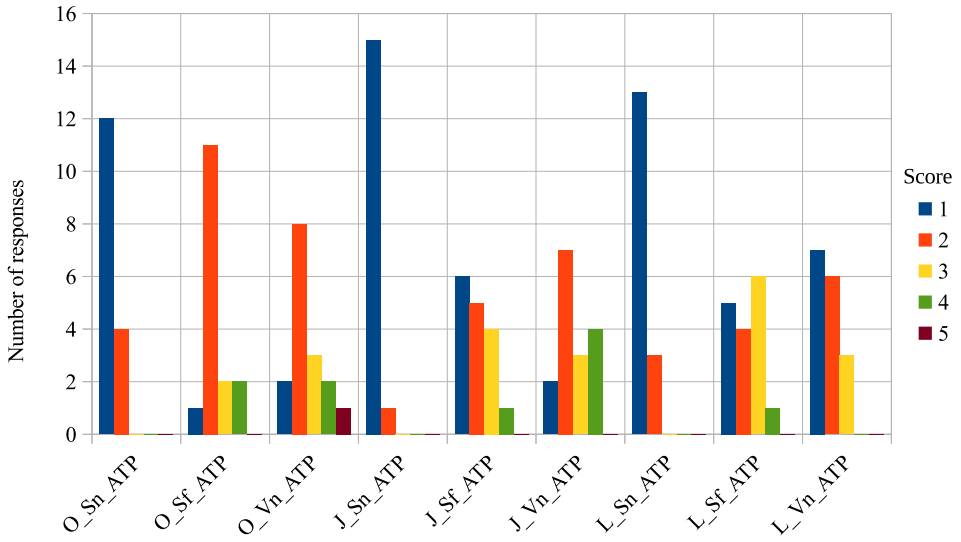

FIG. 18. Responses for various waveforms with full (three-parameter) modulation.

TABLE 6. A summary of ratings for various waveforms with pitch modulation only and with full (three-parameter) modulation.

Sample	Mean score	Median	Sample	Mean score	Median
O_Sn_P	1.13 ±0.33	1.00	O_Sn_ATP	1.25 ±0.43	1.00
O_Sf_P	2.44 ±0.79	2.00	O_Sf_ATP	2.31 ±0.77	2.00
O_Vn_P	2.38 ±0.93	2.00	O_Vn_ATP	2.50 ±1.06	2.00
J_Sn_P	1.13 ±0.33	1.00	J_Sn_ATP	1.06 ±0.24	1.00
J_Sf_P	2.38 ±0.70	2.00	J_Sf_ATP	2.00 ±0.94	2.00
J_Vn_P	2.25 ±0.66	2.00	J_Vn_ATP	2.56 ±1.00	2.00
L_Sn_P	1.00 ±0.00	1.00	L_Sn_ATP	1.19 ±0.39	1.00
L_Sf_P	2.44 ±0.86	2.00	L_Sf_ATP	2.19 ±0.95	2.00
L_Vn_P	2.69 ±1.21	3.00	L_Vn_ATP	1.75 ±0.75	2.00

The fact that sound samples were not level-matched in loudness might have affected the results, some samples were more pronounced than others. However, post-recording level-matching might have affected the results as well, because the level differences were caused by various combinations of temporal and spectral signal modifications, which are perceived differently. A more in-depth study is needed to properly address this issue.

4. Conclusions

A common way to improve the sound produced by a synthesizer leads through enhancing or refining its synthesis algorithm. However, current synthesis algorithms are products of a long way of improvements, and further refinements are difficult. This paper proposed a different approach that can be used in real-time performances. A study was carried out to determine the impact of various types of modulation on the perceived naturalness of the violin sound. Modulation through an automatic low-frequency oscillator was compared with expressive modulation caused by a human using a controller. Two advanced controllers were studied to determine whether simultaneous modulation of more than one parameter can cause a synthetic sound to be perceived as more natural. This has a potential to improve the sound quality of existing synthesizers by changing or refining their controllers.

A synthesizer with an architecture open for controller-related modifications has been designed and used to prepare a set of sound samples, where different waveforms were combined with various modulation sources and modulated parameters. The set was supplemented with real violin recordings. The effect was assessed by a group of expert listeners. The results show a relatively complex situation that requires further study. It has been observed that expressive multi-parameter modulation with advanced controllers brings benefits for waveforms with realistic spectra, close to that of a violin. However, in less realistic waveforms, this kind of modulation may be perceived as less natural than a simple one, such as that obtained through an oscillator.

The conclusion drawn from the study is that, unlike the case of simple MIDI controllers which can be successfully used with their default settings, advanced controllers bring both improvements and problems that need to be addressed. Firstly, one finger cannot effectively control three parameters independently. Both studied controllers handled well control of two parameters, but the third could not have been precisely adjusted, or its range was severely limited. It is partially a design limitation, but further study with other than default mappings between parameters and controller degrees of freedom may bring a solution. Secondly, an expression must be coherent with a signal. If the signal features differ too much from the original instrument, the overly expressive control characteristic of this instrument is perceived as unnatural. Finally, continuous multi-parameter control closes the gap between real instruments, such as the violin, and synthesizers. This implies that performance with such synthesizers will be much more demanding than with simple MIDI devices. They would require a much more thorough learning and training process on the side of the performer. However, as the results show, the perception of expressive sounds clearly improves the synthesis effect if the process is implemented properly, which gives an impulse for future studies.

Fundings

This research was funded by the Department of Mechanics and Vibroacoustics of AGH University of Krakow, Poland, grant no. 10 000-501.00-130 000.

Conflict of interest

The author declares that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Authors' contribution

Marek Pluta conceptualized the study, wrote the original draft, wrote computer programs used within the study, prepared sound samples, performed data interpretation and analysis, reviewed and approved the final manuscript.

References

1. Donati E., Chousidis C. (2022), Electroglottography based real-time voice-to-MIDI controller, *Neuroscience Informatics*, **2**(2): 100041, https://doi.org/10.1016/j.neuri.2022.100041.
2. Fritz C., Stoppani G., Igartua U., Woodhouse J. (2025), Developing methodologies to study perceived sound qualities of violins, *Acta Acustica*, **9**: 32, https://doi.org/10.1051/aacus/2025014.
3. Gurevich M., von Muehlen S. (2001), The accordiatron: A MIDI controller for interactive music, [in:] *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 27–29, https://doi.org/10.5281/zenodo.1176364.
4. Hawley S.H., Chatziioannou V., Morrison A. (2020), Synthesis of musical instrument sounds: Physics-based modeling or machine learning?, *Acoustics Today*, **16**(1): 20–28, https://doi.org/10.1121/AT.2020.16.1.20.
5. Iverson P., Krumhansl C.L. (1993), Isolating the dynamic attributes of musical timbre, *Journal of Acoustical Society of America*, **94**(5): 2593–2603, https://doi.org/10.1121/1.407371.
6. Kim D., Dong H.-W., Jeong D. (2025), ViolinDiff: Enhancing expressive violin synthesis with pitch bend conditioning, [in:] *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, https://doi.org/10.1109/ICASSP49660.2025.10890613.
7. Liu Q. (2024), An FM-wavetable-synthesized violin with natural vibrato and bow pressure, [in:] *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, pp. 243–250, https://doi.org/10.2991/978-94-6463-370-2_27.
8. McAdams S., Bruno G.L. (2012), The perception of musical timbre, [in:] *Oxford Handbook of Music Psychology*, Hallam S., Cross I., Thaut M.H. [Eds], pp. 72–80, Oxford Library of Psychology, https://doi.org/10.1093/oxfordhb/9780199298457.013.0007.
9. Pérez Carrillo A.A. (2009), *Enhancing spectral synthesis techniques with performance gestures using the violin as a case study*, Ph.D. Thesis, Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona.
10. Robertson A. (2011), Seaboard: A new piano keyboard-related interface combining discrete and continuous control, [in:] *Proceedings of the International Conference on New Interfaces for Musical Expression*, https://doi.org/10.5281/zenodo.1178081.
11. Schoonderwaldt E., Friberg A. (2001), Towards a rule-based model for violin vibrato, [in:] *Proceedings of the Workshop on Current Research Directions in Computer Music*, pp. 61–64.
12. Wu Y. *et al.* (2022), MIDI-DDSP: Detailed control of musical performance via hierarchical modeling, [in:] *International Conference on Learning Representations (ICLR)*, pp. 1–27.

Archives of Acoustics Vol. 51, No. 2, pp. 415–426 (2026), https://doi.org/10.24423/archacoust.2026.4332

OSA 2025

Impact of Perforated Sheet Geometry on the Insertion Loss of Absorptive Silencers

Kamil Wójciak*, Joanna Maria Kopania, Patryk Gaj

Institute of Power Engineering – National Research Institute
Warsaw, Poland

*Corresponding Author: kamil.wojciak@itc.edu.pl

Received September 12, 2025; revised December 15, 2025; accepted March 18, 2026; available online March 26, 2026; version of record June 3, 2026; published issue June 24, 2026.

This study evaluates the influence of perforated sheet geometry on the acoustic and aerodynamic performance of absorptive silencers. A modular silencer is developed, enabling the installation of six different perforated metal sheets with varying hole shapes (round, square, elongated), sizes (2 mm to 20 mm), and open area ratios (22 % to 45 %). Glass wool is used as the sound-absorbing filling. Insertion loss, self-noise, and pressure drop are measured in a large reverberation chamber, within the frequency range from 50 Hz to 10 000 Hz, for airflow velocities of 4 m/s, 6 m/s, and 8 m/s. The results indicate that all configurations provide comparable attenuation at low frequencies. Silencers with small round perforations (diameter 2 mm to 6 mm) ensure higher insertion loss and lower self-noise in the mid-frequency range of 1000 Hz to 5000 Hz, without any measurable increase in pressure drop compared to variants with larger or elongated holes. For frequencies above 6300 Hz, perforated sheets with larger holes perform better. Pressure loss differences between all configurations do not exceed 1 Pa at a given flow velocity. The results confirm that aperture size is the primary parameter affecting silencer acoustic effectiveness, while aperture shape and perforation ratio are secondary. These findings provide practical guidelines for optimal silencer design in ventilation systems, ensuring maximum noise reduction with minimal airflow resistance.

Keywords: silencers, insertion loss, experimental test, aeroacoustics, ventilation.

Notations

D_i – insertion loss,
L_W – sound power level,
L_{WI} – sound power level with the test object installed,
L_{WII} – sound power level with the substitution element replacing the test object,
v – flow velocity,
Δp – pressure loss.

1. Introduction

Noise generated by mechanical ventilation systems has become a significant challenge in residential buildings, directly affecting occupant comfort and sleep quality (Harvie-Clark *et al.*, 2019; Abramkina, 2023). Excessive sound levels often lead residents to disable or reduce the use of ventilation, resulting in poor indoor air quality and potential health risks, especially in modern airtight dwellings (Harvie-Clark *et al.*, 2019; Lan *et al.*, 2021). Studies confirm that, while adequate ventilation is essential for maintaining healthy living conditions and improving sleep, these benefits are only realised if ventilation noise is kept to a minimum (Lan *et al.*, 2021). Furthermore, official standards and guideline values for acceptable noise levels in living spaces vary across countries, but recent

recommendations highlight the need to maintain bedroom noise levels below 30 dB(A) to prevent adverse effects on sleep and well-being (Harvie-Clark *et al.*, 2019). Noise in ventilation systems originates from multiple sources such as fans, dampers, control devices, and turbulent airflow, and the overall perceived noise is often exacerbated by improper design or installation (Abramkina, 2023). Consequently, the application of effective acoustic silencers within ventilation systems is crucial for reducing noise emissions, as demonstrated by both engineering analyses and practical experience (Abramkina, 2023).

Perforated metal sheets are widely used as interface elements in absorptive silencers to separate the sound-absorbing material from the duct interior while permitting airflow. The acoustic performance of such silencers is influenced by the properties of the perforated panel, including perforation geometry, open area ratio, and the physical dimensions of both the panel and the absorber. Previous studies have primarily focused on micro-perforated panels (MPPs), where sub-millimetre holes provide the necessary acoustic resistance to achieve broadband sound absorption without the use of porous materials (Maa, 1998; Wu, 1997). The theoretical basis for MPPs, as outlined by Maa (1998), models the holes as arrays of short tubes, characterising the system in terms of acoustic impedance, porosity, and perforation constant, which collectively determine the absorption bandwidth and resonance frequency. MPPs have been implemented in a range of noise control applications, such as room acoustics (Kang, Brocklesby, 2005; Hoshi *et al.*, 2020), ventilation systems (Yu *et al.*, 2016; Zhang *et al.*, 2020; Wójciak *et al.*, 2025), and transparent noise barriers (Asdrubali, Pispola, 2007).

Despite the extensive literature on micro-perforated solutions, the majority of these studies are limited to circular holes with diameters below 1 mm, and relatively little attention has been paid to panels with larger or non-circular perforations in silencer configurations. Classic MPP models do not account for the flow and acoustic effects associated with larger-scale or differently shaped openings. In practice, many industrial silencers use panels with millimetre-scale and non-circular perforations, where the contribution to both acoustic attenuation and flow characteristics may differ fundamentally from micro-perforated designs (Wu, 1997; Allam, Ĺbom, 2011).

The acoustic interaction between the perforated sheet, the porous absorber, and the silencer geometry is also affected by backing cavity configuration, flow presence, and the spatial coupling of acoustic modes (Yang *et al.*, 2015; Yang, Cheng, 2016). For typical absorptive silencers utilising porous materials, increasing the open area ratio and altering perforation geometry modify the sound transmission and pressure loss characteristics, as well as the generation of self-noise (Maa, 1998). However, the impact of non-micro-perforated panels (particularly those with varying hole shapes) on silencer efficiency, self-generated noise, and flow resistance has seldom been assessed.

Acoustic silencers play a pivotal role in reducing ventilation noise, and their design and selection must be tailored to the specific acoustic requirements and airflow conditions of each system. However, there is a lack of comprehensive data on how perforated sheet geometry in absorptive silencers affects noise attenuation, self-noise, and pressure loss. The present study investigates the effect of perforated metal sheets with different hole sizes and geometries, beyond the micro-perforation range, on the effectiveness of absorptive silencers. Using a modular silencer test rig with interchangeable side panels, the study enables direct comparison of circular, square, and elongated perforations with different open area ratios. Measurements were performed in accordance with ISO 7235 (International Organization for Standardization [ISO], 2003), using glass wool as the absorptive filling, and insertion loss was evaluated over a broad frequency range in third-octave bands. The outcomes provide insight into the relationship between perforated sheet parameters and silencer performance under practical conditions, contributing to the optimisation of silencer design in industrial applications.

2. Test objects

The subject of the investigation is an absorptive silencer designed to assess the influence of perforated sheet properties on insertion loss, flow noise and pressure loss. In this configuration, the perforated sheet is utilised as a partition between the airflow channel within the silencer and the adjacent layer of sound-absorbing material. The silencer features a modular construction, comprising an outer casing equipped with two interchangeable side panels, thus enabling the installation of panels with a maximum thickness of 300 mm. The remaining two sides

of the enclosure are fabricated from 3 mm-thick steel sheet, providing mechanical rigidity and acoustic isolation. The 3D model of the silencer constructed for measurements is presented in Fig. 1.

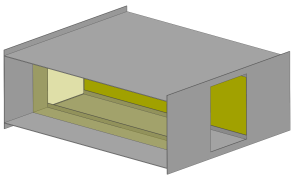

FIG. 1. 3D model of the silencer constructed for measurements.

The silencer is fitted with a rectangular connecting flange having internal dimensions of 315 mm by 250 mm. The total length of the silencer, measured between the inlet and outlet connections, is 1000 mm. The replaceable side panels have dimensions of 320 mm by 1000 mm, allowing for direct comparison of the acoustic effectiveness of various perforated sheets and reference materials. The principal dimensions of the silencer are shown in Fig. 2. The sound-absorbing material selected for the study is Ventilux 6335 glass wool, with a thickness of 100 mm and a density of 35 kg/m^3 (GAJ *et al.*, 2019), which is positioned directly behind the perforated sheet along the airflow path. The same set of glass wool slabs is used for all perforated-sheet variants and all measurement series, ensuring identical acoustic filling conditions for each test configuration.

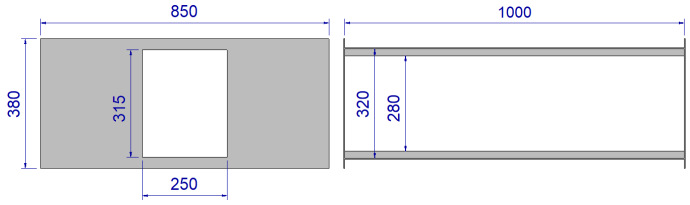

FIG. 2. Principal dimensions of the silencer (in millimetres).

Six types of perforated sheets are evaluated, with systematic variation in both hole geometry and open area ratio. Perforated sheets designated Rv18-26, Qv12-18, and Lv20×5-25×17 are manufactured from 1.5 mm-thick steel sheet, whereas perforated sheets Rv6-12, Rv4-8, and Rv2-3 are manufactured from 1.0 mm-thick steel sheet. Schematic representations of the perforated sheet samples used in the investigation are shown in Fig. 3. The geometric parameters of all tested perforated sheets are summarised in Table 1.

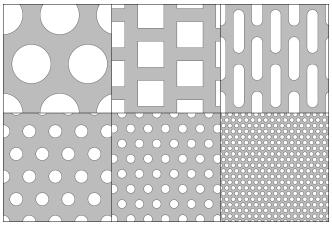

FIG. 3. Schematic views of the perforated sheet samples Rv18-26, Qv12-18, Lv20×5-25×17, Rv6-12, Rv4-8, and Rv2-3 used as side panels in the tested silencer.

TABLE 1. Geometric parameters of the perforated sheets used as inner liners in the absorption silencer.

Designation	Hole shape	Hole size	Pattern	Open area [%]
Rv18-26	Round	18 mm diameter	60° staggered	43.6
Qv12-18	Square	12 mm side		44.0
Lv20×5-25×17	Elongated	20 mm × 5 mm		44.5
Rv6-12	Round	6 mm diameter		22.8
Rv4-8	Round	4 mm diameter		22.8
Rv2-3	Round	2 mm diameter		40.4

For additional comparison, insertion loss measurements are carried out for a configuration comprising only the glass wool, without a perforated sheet.

3. Experiment

3.1. Test rig and operating conditions

The measurements are conducted on a test rig constructed in accordance with ISO 7235 standard (ISO, 2003). The reverberation chamber has a volume of 237.0 m^3 and an area of 231.5 m^2, with non-parallel, sound-reflecting walls. During the measurement of self-noise and pressure drop, the tested silencer is connected to a centrifugal fan via two absorption silencers (Fig. 4). During the measurement of insertion loss, the tested silencer is connected to an external noise source (Fig. 5). The fan, noise source, and tested silencer are positioned outside the reverberation chamber, whereas the outlet is situated inside the chamber. The silencer remains fixed in its position throughout the entire test. Only the upstream source system is reconfigured once, by replacing the fan and its silencers with the loudspeaker source used for the insertion loss measurements. For each perforated-sheet variant, the side panels of the silencer are installed twice: first for the self-noise and pressure-drop measurements, and subsequently for the insertion loss measurements, using the same glass wool slabs in both assemblies.

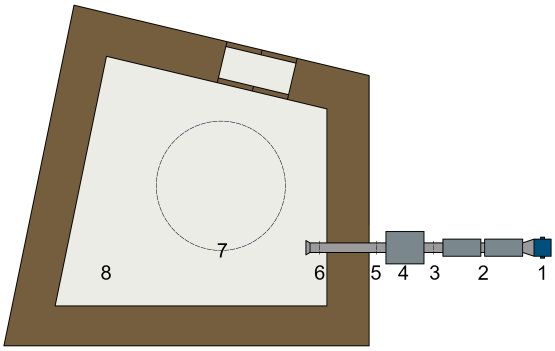

Fig. 4. Test stand with reverberation room scheme for self-noise and pressure drop measurements: 1) fan, 2) set of two silencers, 3) pressure and temperature measurement, 4) test object, 5) pressure measurement, 6) flow velocity measurement, 7) microphone path, 8) reverberation room.

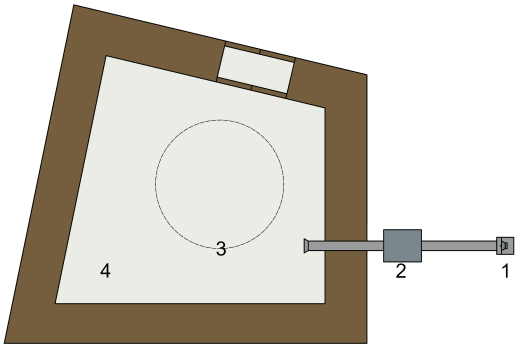

Fig. 5. Test stand with reverberation room scheme for insertion loss measurements: 1) noise source, 2) test object, 3) microphone path, 4) reverberation room.

The airflow rate is controlled by adjusting the frequency of the fan motor using a three-phase inverter, which enables tests at various inlet velocities (4 m/s, 6 m/s, and 8 m/s). The volumetric flow is determined using a Prandtl tube in accordance with ISO 5221 standard (ISO, 1984), with measurements performed in the inlet duct. The static pressure drop across the tested element is measured both upstream and downstream at four circumferentially spaced points using an electronic differential pressure transducer.

3.2. Acoustic and aerodynamic measurements

Acoustic measurements focus on the determination of sound power level in accordance with ISO 3741 standard (ISO, 2010), which specifies precision methods for reverberation chambers. The measuring system comprises a class 1 Norsonic Nor140 sound analyser, a Nor265 rotary table, and the Nor850 software suite. Sound

pressure levels are recorded at twelve points distributed uniformly along a circle with a radius of 1.7 m (circumference 10.7 m) inside the chamber. For each operating condition and each configuration, a single measurement sequence is performed, during which the microphone is successively placed at all twelve positions, while the silencer and duct system remain unchanged. Measurements are performed in 1/3-octave bands within the frequency range from 50 Hz to 10 000 Hz, with each acquisition lasting 30 s.

Background noise is recorded for each measurement series with airflow switched off, enabling calculation of the background correction. Reverberation time is measured for four omnidirectional loudspeaker positions with three microphone positions. All sound power level calculations are completed using a dedicated calculation sheet.

Prior to and following each measurement sequence, instrument calibration is conducted using a class 1 Norsonic Nor1256 calibrator. Additionally, after each measurement set, temperature, relative humidity, and atmospheric pressure are measured and documented, ensuring accurate correction and traceability in the sound power calculation process.

3.3. Statistical analysis

In order to verify whether the differences between the measured values are statistically significant at the 5 % significance level, Student's t-test for two means under the assumption of equal population variances is applied. The statistical procedure follows the approach described in (Montgomery, Runger, 2018). Two independent samples with sizes n_1 and n_2 are considered, for which the sample means $\overline{x_1}$, $\overline{x_2}$, and standard deviations s_1, s_2 are determined. The sample standard deviations are similar in magnitude, which justifies the assumption of homogeneity of variances.

The null hypothesis states that the mean values of the investigated quantity (e.g., insertion loss) under the two measurement conditions are equal:

$$H_0 : \mu_1 = \mu_2, \tag{1}$$

with the alternative hypothesis:

$$H_1 : \mu_1 \neq \mu_2, \tag{2}$$

corresponding to a two-sided test at the significance level $\alpha = 0.05$. In the first step, the pooled estimate of the common variance is computed as

$$s_p^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}, \tag{3}$$

and the test statistic is then given by

$$t_0 = \frac{\overline{x_1} - \overline{x_2}}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \tag{4}$$

The statistic t_0 follows a Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom. Based on the calculated value of t_0 and the degrees of freedom $df = n_1 + n_2 - 2$, the p-value is obtained or the critical region determined using the quantile $t_{\alpha/2,df}$. The null hypothesis H_0 is rejected if

$$|t_0| > t_{\alpha/2,df}, \tag{5}$$

which indicates that the difference between the sample means is statistically significant at the 95 % confidence level. Otherwise, there is no statistical evidence to conclude that the mean values differ between the compared silencer configurations.

4. Results

4.1. Insertion loss

Insertion loss is applied to quantify the acoustic performance of the tested silencers. Insertion loss D_i is defined as the reduction in sound power level measured downstream of the silencer resulting from the replacement

of a reference duct section with the silencer under test. In the present study, the reference section is a silencer with glass wool panels covered on the inner side by a solid (non-perforated) metal sheet, whereas the test objects are silencers with glass wool panels covered by a perforated metal sheet or, in one variant, with exposed glass wool. Insertion loss is calculated according to:

$$D_i = L_{WII} - L_{WI}, \tag{6}$$

where L_{WI} is the sound power level in the considered frequency band measured downstream of the silencer with the perforated sheet (or exposed wool), and L_{WII} is the sound power level in the same frequency band measured downstream of the reference silencer with a solid (non-perforated) metal sheet.

This approach enables the evaluation of the acoustic effectiveness associated with the use of perforated sheets of varying aperture size and shape. In addition, a single-number A-weighted insertion loss value is determined for a flat-spectrum noise, assuming a constant sound power level in all frequency bands prior to A-weighting.

Table 2 presents the calculated insertion loss values in octave bands and the single-number A-weighted values for silencers fitted with various internal perforated metal sheets, as well as for the configuration with exposed glass wool. The insertion loss spectra in 1/3-octave bands for all tested configurations are shown in Fig. 6.

Table 2. Insertion loss values in octave bands for silencers with various internal perforated sheets.

f [Hz]	D_i [dB]						
	Rv18-26	Qv12-18	Lv20×5-25×17	Rv6-12	Rv4-8	Rv2-3	Wool itself
63	0.2	−0.1	−0.2	−0.2	−0.5	−0.5	0.0
125	0.9	1.0	0.6	0.6	0.5	0.5	0.8
250	4.2	4.6	4.5	4.9	4.9	4.8	5.0
500	8.5	8.4	8.3	8.4	8.6	8.3	8.3
1000	13.0	13.2	13.4	13.6	13.6	13.3	13.6
2000	6.5	6.6	6.5	8.1	8.2	8.1	8.6
4000	3.5	3.7	3.6	4.7	4.7	4.5	4.7
8000	1.5	1.4	1.4	0.9	0.8	0.8	0.4
A	5.2	5.3	5.2	5.6	5.6	5.5	5.4

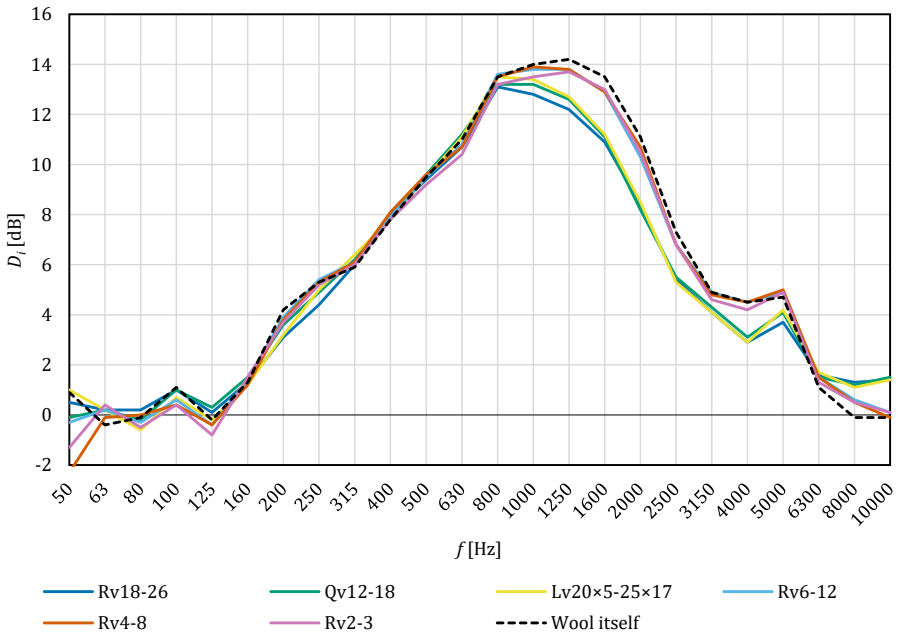

Fig. 6. Insertion loss spectra in 1/3-octave bands for silencers with different internal perforations.

Notable differences are observed between the tested perforated sheets, particularly in the mid- and high-frequency ranges.

In the lowest octave bands (up to 125 Hz), insertion loss is minimal and does not exceed 1 dB for any variant, indicating limited low-frequency attenuation regardless of perforation type. In the range from 250 Hz to 1000 Hz, all configurations achieve similar insertion loss, with values peaking in the 800 Hz to 1250 Hz band.

Above 1000 Hz, the insertion-loss spectra of the individual silencers start to diverge. In the range 1000 Hz to 5000 Hz, silencers with smaller round apertures (Rv6-12, Rv4-8, Rv2-3) generally exhibit higher insertion loss than variants with larger apertures (Rv18-26, Qv12-18, Lv20×5-25×17). At the highest frequencies (above 6300 Hz), the trend reverses, and perforated sheets with larger holes tend to provide higher insertion loss.

To quantify these differences, a two-sample Student's t-test with pooled variance was applied. The analysis shows that the differences between the groups with smaller and larger apertures are statistically significant at the 5 % significance level in the frequency range from 1250 Hz to 2500 Hz and at 4000 Hz and 10 000 Hz, whereas at the remaining frequencies no statistically significant differences between the compared groups are confirmed.

These findings indicate that the type of internal perforated sheet influences insertion loss primarily in the mid- to-high frequency range. Perforated sheets with smaller apertures provide more effective attenuation in those bands where statistically significant differences are observed (mainly from 1250 Hz to 2500 Hz and around 4000 Hz), while at other frequencies the apparent trends in the spectra are not supported by statistical evidence. The shape of the aperture has a secondary effect relative to aperture size.

4.2. Self-noise

Table 3, Table 4, and Table 5 present octave-band sound power levels and overall A-weighted values reported for the self-noise of all tested silencer configurations: perforated sheets, glass wool without an internal sheet, and solid sheet. Each table contains results obtained for a single flow velocity, namely 4 m/s, 6 m/s or 8 m/s. Figure 7 to Fig. 9 illustrate the self-noise spectra in 1/3-octave bands for all tested silencers at flow velocities of 4 m/s, 6 m/s, and 8 m/s, respectively.

Table 3. Sound power levels of self-noise for all silencer configurations (perforated sheets, glass wool, solid sheet): octave-band values and overall A-weighted value at a flow velocity of 4 m/s.

f [Hz]	L_W [dB] at flow velocity 4 m/s							
	Rv18-26	Qv12-18	Lv20×5-25×17	Rv6-12	Rv4-8	Rv2-3	Wool itself	Sheet metal
63	45.0	46.6	45.3	45.3	44.8	45.0	45.0	46.6
125	34.1	34.5	34.1	34.0	34.8	34.7	34.9	35.7
250	34.1	34.5	33.9	33.0	32.7	32.8	32.5	39.7
500	28.3	28.8	28.4	28.0	27.7	27.8	27.7	37.3
1000	18.5	19.1	18.6	17.9	17.7	17.8	17.6	29.5
2000	15.8	16.0	15.9	15.1	15.0	14.9	14.8	21.4
4000	17.3	17.7	17.4	16.8	16.8	16.7	16.7	18.2
8000	20.3	21.1	20.3	20.1	20.1	20.0	20.1	20.5
A	30.2	30.7	30.2	29.6	29.4	29.4	29.3	37.1

Table 4. Sound power levels of self-noise for all silencer configurations (perforated sheets, glass wool, solid sheet): octave-band values and overall A-weighted value at a flow velocity of 6 m/s.

f [Hz]	L_W [dB] at flow velocity 6 m/s							
	Rv18-26	Qv12-18	Lv20×5-25×17	Rv6-12	Rv4-8	Rv2-3	Wool itself	Sheet metal
63	53.2	55.2	53.6	52.9	52.6	52.8	53.1	55.0
125	42.7	42.7	42.7	42.6	42.7	42.7	42.6	44.0
250	42.6	42.7	42.4	41.1	41.2	41.2	40.6	47.6
500	37.4	37.7	37.2	36.9	36.9	36.3	36.1	45.6
1000	31.1	31.5	31.0	30.3	30.2	30.1	29.5	42.2
2000	26.6	27.2	26.8	25.8	25.5	25.2	24.4	37.3
4000	24.6	25.2	24.9	24.1	23.8	23.7	23.2	30.4
8000	21.1	21.9	21.1	20.9	20.9	20.9	20.9	22.2
A	39.1	39.5	39.1	38.3	38.3	38.0	37.6	47.1

TABLE 5. Sound power levels of self-noise for all silencer configurations (perforated sheets, glass wool, solid sheet): octave-band values and overall A-weighted value at a flow velocity of 8 m/s.

f [Hz]	L_W [dB] at flow velocity 8 m/s							
	Rv18-26	Qv12-18	Lv20×5-25×17	Rv6-12	Rv4-8	Rv2-3	Wool itself	Sheet metal
63	58.6	59.5	59.1	57.5	56.9	57.0	57.8	58.7
125	49.0	48.8	49.1	48.5	48.8	48.4	48.6	50.0
250	48.6	48.7	48.4	46.9	47.2	47.0	46.7	53.4
500	43.6	44.0	43.5	42.8	43.0	42.3	42.0	51.2
1000	39.0	39.4	39.0	38.3	38.3	37.6	37.0	49.2
2000	36.5	37.0	36.7	35.7	35.2	35.2	34.3	46.6
4000	34.4	34.9	34.8	33.9	33.6	33.4	33.0	40.9
8000	26.7	27.2	26.9	26.6	26.5	26.4	26.1	30.6
A	46.2	46.5	46.2	45.2	45.3	44.8	44.4	54.1

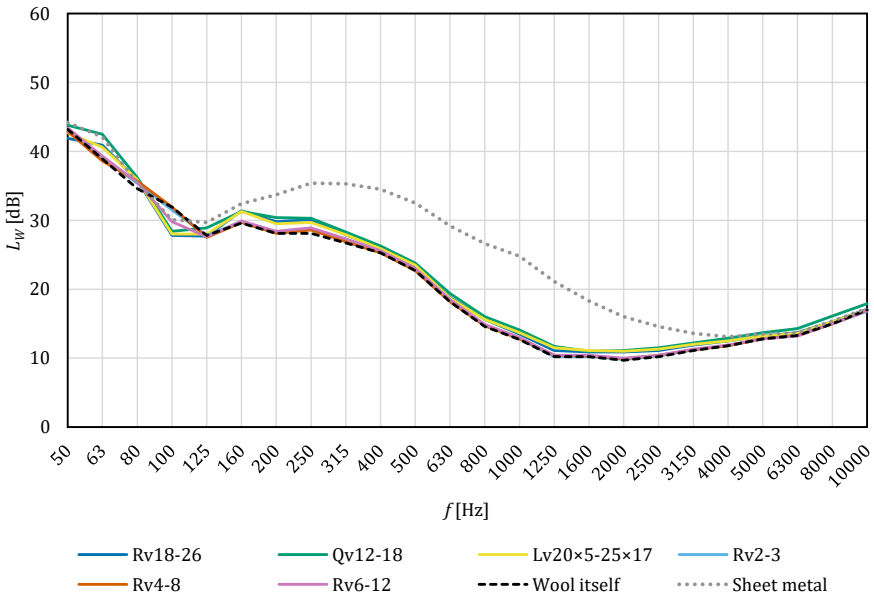

FIG. 7. Self-noise spectra (1/3-octave bands) of silencers at 4 m/s flow velocity.

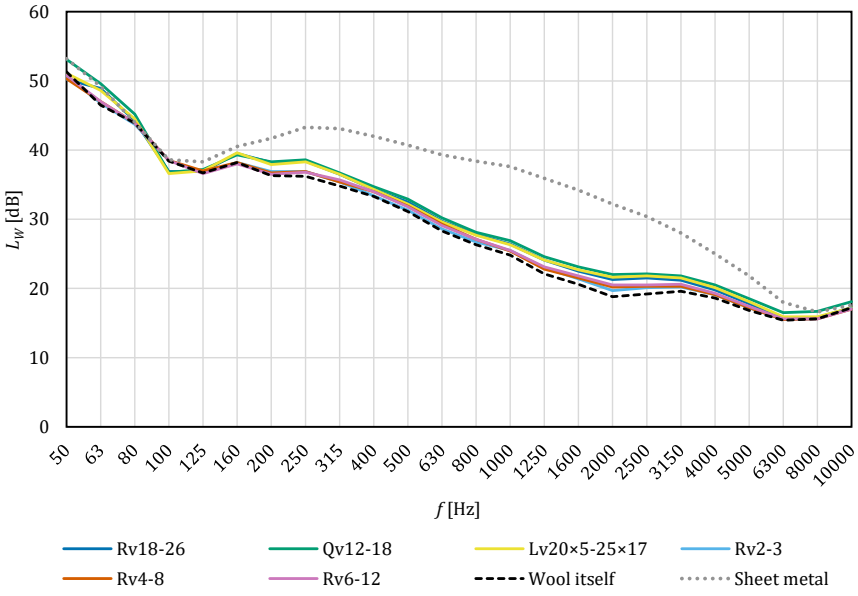

FIG. 8. Self-noise spectra (1/3-octave bands) of silencers at 6 m/s flow velocity.

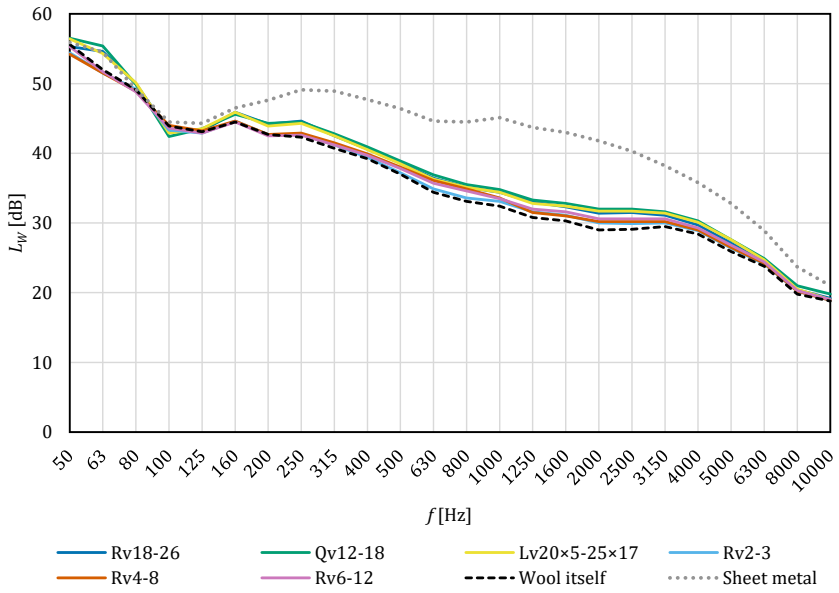

FIG. 9. Self-noise spectra (1/3-octave bands) of silencers at 8 m/s flow velocity.

At a flow velocity of 4 m/s, the self-noise levels of all silencers with perforated sheets are similar and remain within a narrow range. The configuration containing only glass wool exhibits the lowest overall self-noise. In the lowest frequency bands (63 Hz to 250 Hz), differences between perforated sheet variants are negligible, and the levels are close to those observed for this configuration.

To quantify the observed differences between perforated sheet types, a two-sample Student's t-test was applied. For 4 m/s, the analysis indicates that the differences in self-noise between silencers with small apertures (Rv6-12, Rv4-8, Rv2-3) and those with larger or elongated apertures (Rv18-26, Qv12-18, Lv20×5-25×17) are statistically significant at the 5 % significance level in the 160 Hz to 315 Hz and 800 Hz to 4000 Hz bands, whereas at the remaining frequencies no statistically significant differences between these two groups are confirmed.

As the flow velocity increases to 6 m/s and 8 m/s, self-noise levels rise for all configurations. At both higher velocities, silencers with the smallest apertures (Rv6-12, Rv4-8, Rv2-3) generally display lower self-noise values than those with larger or elongated apertures (Rv18-26, Qv12-18, Lv20×5-25×17), particularly in the 500 Hz to 4000 Hz range. For 6 m/s, the t-test shows statistically significant differences between these two groups in the 160 Hz to 315 Hz and 800 Hz to 2500 Hz bands, while for 8 m/s the statistically significant range extends from 160 Hz to 500 Hz and 800 Hz to 2500 Hz. Outside these bands, the apparent differences in the spectra are not confirmed as statistically significant at the 5 % level. The configuration with a solid (non-perforated) inner sheet consistently generates the highest self-noise levels across all velocities, especially in the mid-frequency range, whereas the configuration with exposed wool maintains comparatively low self-noise at all flow rates.

A detailed analysis of the spectral distributions indicates that the largest differences between perforated sheet variants occur in the 160 Hz to 5000 Hz range, particularly at 8 m/s. At the highest flow velocity, silencers with larger apertures tend to exhibit increased self-noise in this range compared to those with small apertures.

In summary, the type of perforation in the internal sheet of the silencer has a measurable influence on self-noise, especially at elevated flow velocities and in the 160 Hz to 5000 Hz range. Perforated sheets with smaller apertures provide a moderate reduction in self-noise relative to variants with larger or elongated apertures in those bands where statistically significant differences are observed, while a solid sheet as the inner lining results in unfavourable self-noise performance over most of the spectrum.

4.3. Pressure loss

Table 6 presents the measured pressure loss values for the silencer with various internal perforated sheets, as well as for the configurations with exposed glass wool and a solid (non-perforated) sheet, at flow velocities of

Table 6. Pressure loss of silencers with various perforated sheets at different flow velocities.

v [m/s]	Δp [Pa]							
	Rv18-26	Qv12-18	Lv20×5-25×17	Rv6-12	Rv4-8	Rv2-3	Wool itself	Sheet metal
4	1.8	1.8	1.9	1.8	2.0	1.4	1.6	2.1
6	4.0	4.3	3.9	3.8	4.2	3.5	3.8	4.2
8	6.8	7.1	6.3	6.3	7.0	6.2	6.6	7.0

4 m/s, 6 m/s, and 8 m/s. Figure 10 shows the average pressure loss as a function of flow velocity for all tested configurations.

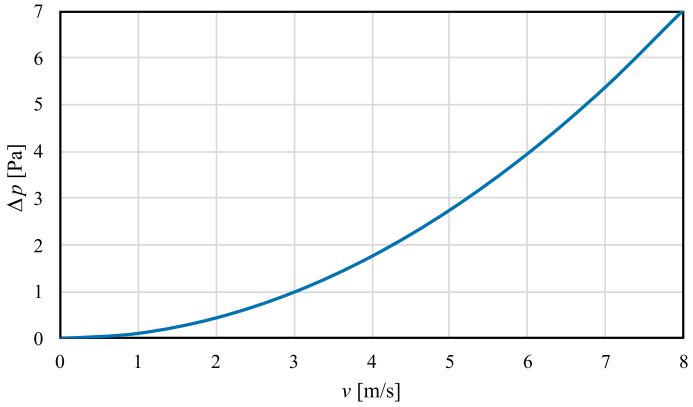

Fig. 10. Average pressure loss of silencers as a function of flow velocity.

The pressure losses observed for all variants increase with flow velocity and remain consistent across different types of internal perforated sheets. For each flow velocity, the differences between configurations do not exceed 1 Pa. Both the configuration with exposed glass wool and that with a solid sheet exhibit pressure losses within the same range as the perforated sheet variants.

These results indicate that the type and geometry of the internal perforated sheet have a negligible effect on the overall pressure loss of the silencer, and the pressure loss characteristics are primarily governed by the total flow resistance of the silencer assembly.

5. Conclusions

This paper investigated the impact of perforated sheet metal on the performance of absorptive silencers, focusing on insertion loss, self-noise, and pressure loss characteristics. The study utilised a modular silencer test rig with replaceable side panels, allowing for the evaluation of various perforated sheet geometries (circular, square, elongated), sizes and perforation ratios. Glass wool (Ventilux 6335, 100 mm thick) served as the sound-absorbing material. Insertion loss measurements were conducted in a reverberation chamber according to ISO 7235 (ISO, 2003), across the 50 Hz to 10 000 Hz frequency range. The same glass wool layers were used in all measurements, ensuring constant properties of the absorption material throughout the test. Furthermore, acoustics measurements were based on averaging over 12 different microphone positions in the reverberation chamber. All acoustic measurements were performed using the same measurement instrumentation.

Insertion loss tests revealed that all silencer variants exhibited minimal attenuation at low frequencies (below 160 Hz). In the 250 Hz to 1000 Hz range, all perforated sheet variants and the wool-only configuration achieved similar, relatively high insertion loss values, with a maximum around 1000 Hz. Above 1000 Hz, silencers with smaller round apertures (Rv6-12, Rv4-8, Rv2-3) provided higher insertion loss in the 1000 Hz to 5000 Hz range, whereas for frequencies above 6300 Hz, perforated sheets with larger apertures (Rv18-26, Qv12-18, Lv20×5-25×17) offered superior attenuation. The statistical analysis based on two-sample Student's t-test with pooled variance confirmed that the differences in insertion loss between the group with larger or elongated apertures (Rv18-26,

Qv12-18, Lv20×5-25×17) and the group with smaller round apertures (Rv6-12, Rv4-8, Rv2-3) are statistically significant at the 5 % significance level in the 1250 Hz to 2500 Hz bands, as well as at 4000 Hz and 10 000 Hz, while at the remaining frequencies the observed differences were not confirmed as statistically significant.

Self-noise measurements showed that, at a flow velocity of 4 m/s, all tested configurations exhibited similar levels, with the wool-only configuration having the lowest overall self-noise. As the flow velocity increased to 6 m/s and 8 m/s, self-noise levels rise for all silencers and the differences between perforated sheet types became more apparent: silencers with the smallest apertures (Rv6-12, Rv4-8, Rv2-3) consistently exhibited slightly lower self-noise, particularly in the 160 Hz to 5000 Hz range. Silencers equipped with a solid (non-perforated) inner sheet generated the highest self-noise levels across all test conditions, especially in the mid-frequency bands. For all investigated flow velocities, the statistical evaluation indicated that the differences in self-noise between silencers with small round apertures and those with larger or elongated apertures were statistically significant in the 160 Hz to 315 Hz and 800 Hz to 2500 Hz bands, whereas outside these ranges the apparent differences in the spectra were not confirmed as statistically significant at the 5 % level.

Pressure loss measurements demonstrated that, for all tested variants, pressure loss increased proportionally with airflow velocity and remained consistent between different perforated sheet types. The differences in pressure drop for a given velocity did not exceed 1 Pa, and the absolute values were similar for perforated, solid, and wool-only configurations. Thus, the geometry of the internal perforated sheet had a negligible effect on pressure loss in this silencer type.

A key observation is that the Rv18-26 and Rv2-3 perforated sheets had identical aperture shape (round) and a similar perforation ratio (43.6 % for Rv18-26 and 40.4 % for Rv2-3), yet their performance differed significantly. The Rv2-3 sheet, with smaller holes, provided much higher insertion loss and lower self-noise in the mid-frequency range compared to Rv18-26, which had much larger apertures. Therefore, it can be concluded that it is primarily the aperture size (not the shape or perforation ratio) that governs acoustic effectiveness in absorptive silencers. This conclusion is supported by the statistical results, which systematically demonstrate significant differences between the groups with small and large apertures in the mid-frequency range for both insertion loss and self-noise. It should be noted that the perforated sheets with smaller apertures (Rv6-12, Rv4-8, Rv2-3) had a thickness of 1.0 mm, whereas the sheets with larger apertures (Rv18-26, Qv12-18, Lv20×5-25×17) were 1.5 mm thick, which may also influence the silencer performance. Although the small- and large-aperture sheets also differed in thickness (1.0 mm vs 1.5 mm), the comparison of Rv2-3 and Rv18-26 (similar open area ratio and identical hole shape) indicates that the aperture diameter is the dominant factor affecting insertion loss.

The results demonstrate that perforated sheets with small round apertures (Rv6-12, Rv4-8, Rv2-3) ensure the most favourable combination of high insertion loss across a wide frequency range and low self-noise, without causing any measurable increase in pressure loss. These findings suggest that, for ventilation and noise control applications, the use of perforated sheets with small circular apertures enables effective noise reduction without compromising airflow performance. In the acoustic measurements, the same test silencer remained in place. The only change in the setup was the sound source, which was switched once from the fan unit with silencers to a loudspeaker system for insertion loss determination. For each perforated configuration, the side panels were mounted twice: once for self-noise and pressure loss measurements, and once for insertion loss measurements after switching the sound source. In both types of acoustic tests (insertion loss and self-noise), panels with smaller perforation (Rv6-12, Rv4-8, Rv2-3) consistently exhibited superior acoustic performance compared to panels with larger apertures.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions

Kamil Wójciak performed the measurements, analysed and interpreted the data, and wrote the original draft. Joanna Maria Kopania conceptualized the study and contributed to the analysis and interpretation of the data. Patryk Gaj performed the measurements and contributed to the analysis and interpretation of the data. All authors reviewed and approved the final manuscript.

References

1. Abramkina D. (2023), Noise from mechanical ventilation systems in residential buildings, *E3S Web of Conferences*, **457**: 02007, https://doi.org/10.1051/e3sconf/202345702007.
2. Allam S., Łbom M. (2011), A new type of muffler based on microperforated tubes, *Journal of Vibration and Acoustics*, **133**(3): 031005, https://doi.org/10.1115/1.4002956.
3. Asdrubali F., Pispola G. (2007), Properties of transparent sound-absorbing panels for use in noise barriers, *The Journal of the Acoustical Society of America*, **121**(1): 214–221, https://doi.org/10.1121/1.2395916.
4. Gaj P., Kopania J., Wójciak K., Bogusławski G. (2019), Assessment of sound absorbing properties of composite made of recycling materials, *Vibrations in Physical Systems*, **30**(1): 2019109.
5. Harvie-Clark J., Conlan N., Wei W., Siddall M. (2019), How loud is too loud? noise from domestic mechanical ventilation systems, *International Journal of Ventilation*, **18**(4): 303–312, https://doi.org/10.1080/14733315.2019.1615217.
6. Hoshi K. *et al.* (2020), Implementation experiment of a honeycomb-backed MPP sound absorber in a meeting room, *Applied Acoustics*, **157**: 107000, https://doi.org/10.1016/j.apacoust.2019.107000.
7. International Organization for Standardization (1984), *Air distribution and air diffusion – Rules to methods of measuring air flow rate in an air handling duct* (ISO Standard No. 5221:1984), https://www.iso.org/standard/11223.html.
8. International Organization for Standardization (2003), *Acoustics – Laboratory measurement procedures for ducted silencers and air-terminal units – Insertion loss, flow noise and total pressure loss* (ISO Standard No. 7235:2003), https://www.iso.org/standard/30385.html.
9. International Organization for Standardization (2010), *Acoustics – Determination of sound power levels and sound energy levels of noise sources using sound pressure – Precision methods for reverberation test rooms* (ISO Standard No. 3741:2010), https://www.iso.org/standard/52053.html.
10. Kang J., Brocklesby M.W. (2005), Feasibility of applying micro-perforated absorbers in acoustic window systems, *Applied Acoustics*, **66**(6): 669–689, https://doi.org/10.1016/j.apacoust.2004.06.011.
11. Lan L., Sun Y., Wyon D.P., Wargocki P. (2021), Pilot study of the effects of ventilation and ventilation noise on sleep quality in the young and elderly, *Indoor Air*, **31**(6): 2226–2238, https://doi.org/10.1111/ina.12861.
12. Maa D.-Y. (1998), Potential of microperforated panel absorber, *The Journal of the Acoustical Society of America*, **104**(5): 2861–2866, https://doi.org/10.1121/1.423870.
13. Montgomery D.C., Runger G.C. (2018), *Applied Statistics and Probability for Engineers*, Wiley, Hoboken.
14. Wu M.Q. (1997), Micro-perforated panels for duct silencing, *Noise Control Engineering Journal*, **45**(2): 69–77, https://doi.org/10.3397/1.2828428.
15. Wójciak K., Kopania J.M., Gaj P. (2025), Acoustic properties of the absorption silencers with a micro-perforated channel in the air-flow, *Vibrations in Physical Systems*, **36**(1): 2025112, https://doi.org/10.21008/j.0860-6897.2025.1.12.
16. Yang C., Cheng L. (2016), Sound absorption of microperforated panels inside compact acoustic enclosures, *Journal of Sound and Vibration*, **360**: 140–155, https://doi.org/10.1016/j.jsv.2015.09.024.
17. Yang C., Cheng L., Hu Z. (2015), Reducing interior noise in a cylinder using micro-perforated panels, *Applied Acoustics*, **95**: 50–56, https://doi.org/10.1016/j.apacoust.2015.02.003.
18. Yu X., Cui F.S., Cheng L. (2016), On the acoustic analysis and optimization of ducted ventilation systems using a sub-structuring approach, *The Journal of the Acoustical Society of America*, **139**(1): 279–289, https://doi.org/10.1121/1.4939785.
19. Zhang X., Yang C., Cheng L., Zhang P. (2020), An experimental investigation on the acoustic properties of micro-perforated panels in a grazing flow, *Applied Acoustics*, **159**: 107119, https://doi.org/10.1016/j.apacoust.2019.107119.