ARCHIVES OF ACOUSTICS
16, 2, 369–381 (1991)

DIRECT APPLICATION OF LINGUISTIC VARIABLE TO THE SPEECH SEGMENT DESCRIPTION IN ISOLATED WORD RECOGNIZER

W. W. WIĘŹLAK

Institute of Fundamental Technological Research Polish Academy of Sciences
(00-049 Warszawa, Świętokrzyska 21)

A simple model of speech recognition based on identification of broad articulatory classes is presented. Direct application of the notion of linguistic variable to description of acoustic parameters evolutions and recognition of broad articulatory classes is discussed. The recognition algorithm is based on a fuzzy automatic network. An example of application of the described method to isolated word recognition is given.

1. Introduction

One of the most essential and important problems in the automatic speech recognition task is that of finding the relations between acoustic signal continuum and the string of phonetic classes which is related to the signal. More precisely, the relations between acoustic cues for the description of certain physical characteristics of the speech signal and phonetic features employed for characterization of speech units, mainly phonemes. Years of research in phonetics and speech physiology have shown that the solution of the "acoustic cues vs. phonetic features" problem is neither simple nor unambiguous and the efforts to established automatic phonetic transcription of the acoustic speech signal have been succesful only for strictly limited conditions. One of the main reasons of such results was that we have to find the complex relations between psychological, i.e. rational rather than physical, units as phonemes, which are defined as the smallest distinctive elements of a word for a particular language, and physical, i.e., measurable parameters like speech spectrum, LPC coefficients etc. Although very sophisticated and precise methods of speech signal measurements and description have been developed, one has to tackle the problem of the descriptive character of phonetic concepts. This vagueness of description results mainly from the approximate definitions of phonetic classes in terms of acoustic cues as pointed above, but another source of vagueness are the coarticulation effects which make the consecutive speech segments influence one another, changing their acoustic characteristics without modifying their phonetic

meanings. In addition, individual and statistical differences across speakers and tokens make the whole task very complicated.

For automatic recognition of isolated or connected words, various strategies of solving this problem have been proposed, from the deterministic pattern recognition methods used by BEZDEL [1], SAMBUR and RABINER [2] or WEINSTEIN and al. [3] to statistical pattern recognition, e.g., [4] and Bayesian classifier [5]. The fuzzy theoretical approach has also attracted the attention of researchers in the last 10 years, offering philosophical concepts and mathematical apparatus which was thought to formalize and overcome the inherent vagueness of the relations between acoustic cues and phonetic features [6], [7].

The study presents how to use the concept of a linguistic variable in describing and recognizing certain phonetic classes in the Isolated Word Recognizer (IWR). Apart from the application of the fuzzy theoretical framework to the system, i.e., direct case of the notion of linguistic variable and fuzzy automata, it is also shown how a careful analysis of a very poor set of parameters, only three used in the present study, supported by an important amount of subjective and objective knowledge incorporated in the recognition algorithm could be used for successful identification of certain broad phonetic classes. General speech description strategy is described elsewhere [7], [8]; however, it consists mainly in detecting and describing certain acoustic events related to the occurrence of broad articulatory phonetic classes.

2. Speech material and parameter extraction

The vocabulary used in this study consists of 60 words (10 digits + 50 command words) provided for a voice controlled minicomputer. Each 60 word was spoken at normal speed by 6 subjects (5 men, 1 woman), two of them with slurred pronunciation. This set made up the learning group. The algorithm was tested on another set of 6 speakers (5 men, 1 woman), 5 of them were new to the system.

The block diagram of IWR is shown in Fig. 1. The incoming speech signal is analysed by the Parametric Speech Analyzer, built at the Speech Acoustic Laboratory IFTR. At the output of the analyzer, a sequence of parameters is sent to the PC microcomputer every 10 ms. Seven parameters were measured:

1. AO – log overall amplitude envelope,
2. LP – log amplitude in low frequency band, 80–800 Hz,
3. HP – log amplitude in high frequency band, 4.5–8 kHz,
4. $F1$, $F2$ – first and second formant frequencies,
5. ZCR – zero-crossing rate,
6. FO – fundamental frequency.

However, only the first three of them AO, LP, HP were used in the study (Fig. 1).

3. The set of speech segments categories

The set of labels used for word description in IWR consisted of 9 names of broad acoustic articulatory classes which resulted from the general classification of

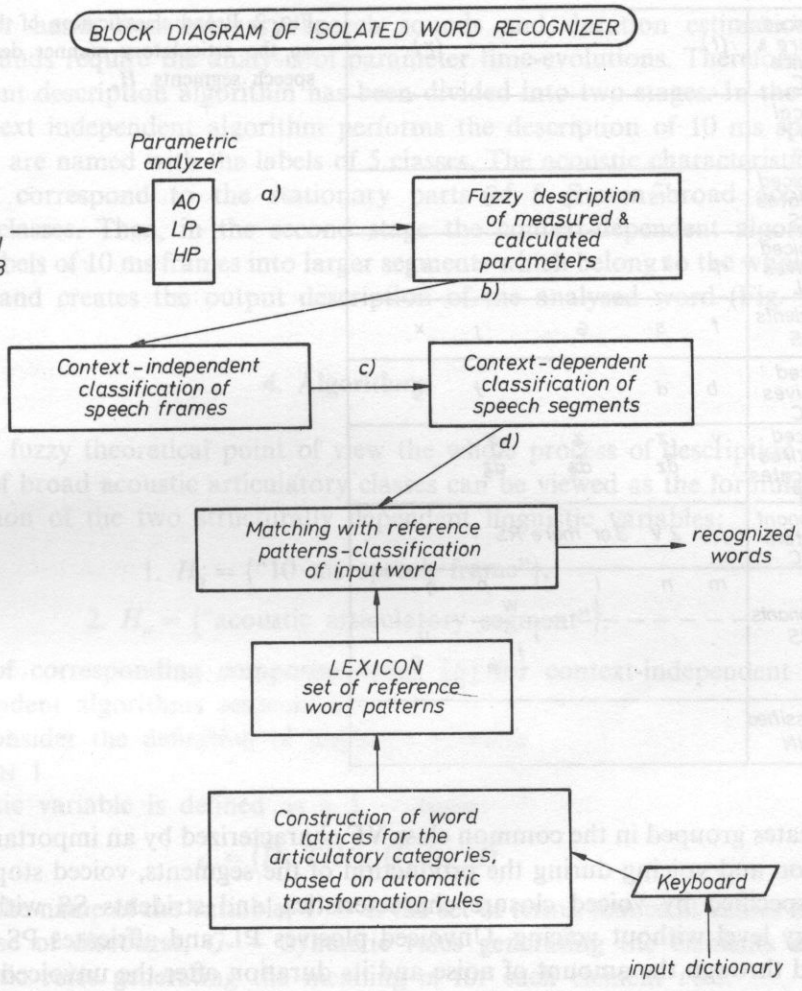

FIG. 1. Block diagram of the isolated word recognizer a) data matrix, b) description of data values with linguistic terms as "low", "medium", "high",... c) string of composite terms describing linguistic variable $H_s = \{10$ ms speech frame$\}$, d) string of composite terms articulatory categories — manner describing linguistic variable $H_a = \{$articulatory segment$\}$.

speech sounds based on the manner of articulation, as it is shown in Fig. 2. Resonants are defined as sounds produced with such a suitable shaping of the vocal tract that the airflow through the mouth and/or nostrils is free. This class contains vowels, nasals, glides and laterals. They are divided into two classes RS and RC on the basis of their duration. All other sounds are called obstruents, made by obstructing the flow of air. The are characterized by an extreme narrowing or constriction at some point of the vocal tract resulting in blocking of the air stream, or in its turbulent character. The definition of obstruents implies various acoustic phenomena to be analysed, within the obstruent part of the speech signal. Therefore they have been divided into 7 classes of segments. There are voiced fricatives and

Unvoiced closure & silence UC	(f) (x)
Apical /r/ AR	r
Unvoiced affricates PS	t͡s t͡ɕ t͡ʂ
Unvoiced plosives PL	p t c k
Stridents SS	f s ɕ ʂ x
Voiced plosives OC	b d ɟ g
Voiced fricatires & affricates VF	v z ʑ ʐ d͡z d͡ʑ d͡ʐ
Resonant clusters RC	2 V 3 or more RS
Resonants RS	m n l ɲ ŋ j w i u ɨ e o a
Unclassified NN	

FIG. 2. Broad classification of the sounds based on the articulatory manner description of the speech segments. H_a

affricates grouped in the common class VF characterized by an important amount of friction and voicing during the production of the segments, voiced stops OC which are specified by voiced closure characteristics, and stridents SS with high noise energy level without voicing. Unvoiced plosives PL and affricates PS are differentiated through the amount of noise and its duration after the unvoiced closure UC. Apical /r/ (AR), the most common realization of the /r/ sound in Polish, presents itself in a different way in the acoustic signal but its salient feature is a sequence of short resonant-obstruent units. In intervocalic position it shows a distinct short dip in the overall amplitude of the signal and before and after resonants it results in saw-tooth-like amplitude variations on the raising or falling edge. So-called unvoiced closure UC is mainly caused by blocking of the airstream but sometimes also by weak aspiration or delay of voicing onset in the sequence of segments with strongly different places of articulation.

Among obstruents one can find two types of sounds: quasi-stationary ones, these which can be produced without restricted timing, e.g., fricatives underlaying the classes SS and VF, and nonstationary speech sounds which without restricted timing loose their phonemic character or cannot be produced, e.g., plosives, /r/ sound. Certainly, stationary segments like SS, UC, partially VF as well as resonants may be identified only through the analysis of instantaneous values of acoustic parameters.

On the other hand, nonstationary speech sounds and duration estimation of stationary sounds require the analysis of parameter time-evolutions. Therefore, the speech segment description algorithm has been divided into two stages. In the first one the context independent algorithm performs the description of 10 ms speech frames which are named with the labels of 5 classes. The acoustic characteristics of these classes correspond to the stationary parts of 5 chosen broad acoustic articulatory classes. Then, in the second stage the context-dependent algorithm gathers the labels of 10 ms frames into larger segments which belong to the whole set of 9 classes and creates the output description of the analysed word (Fig. 1).

4. Algorithms

From the fuzzy theoretical point of view the whole process of description and recognition of broad acoustic articulatory classes can be viewed as the formulation and description of the two structurally dependent linguistic variables:

1. $H_s = \{$"10 ms speech frame"$\}$,
2. $H_a = \{$"acoustic articulatory segment"$\}$,

by the set of corresponding composite terms [6] for context-independent and context-dependent algorithms respectively.

Let us consider the definition of linguistic variable.

DEFINITION 1

A linguistic variable is defined as a 5 – tuple:

$$\Lambda = (H, T(H), U, G, M)$$

where: H ist the name of the variable, T – is the set of terms, names of values of H, U – universe of discourse, G – syntactic rules generating the elements of T, M – semantic rules generating the meaning m for each element $t \in T$.

From the formal definition of linguistic variable given above, in real applications the important question arises how to represent the syntactic and semantic rules which describe the set of terms T. When structural relationship between the terms and variable is not sophisticated, the simplest solution to the problem is to apply fuzzy naming relations describing the elements t in the most straightforward form of syntactic and semantic rules. Moreover, in such a case there exist learning procedures assuring optimal efficiency of the relations in the sense of minimum description error of. [6]. In other cases generative grammars are frequently applied, hence the main drawback of this approach is the hypothesis-and-test structure resulting in the top-down "active" algorithms. Another possibility is to use fuzzy automata (FA) which could make the "passive" network realizations more suitable for bottom-up analysis of acoustic cues. Therefore, taking into account the bottom-up organization of the recognition procedures in IWR, fuzzy naming relations and FA network were chosen for the description of the linguistic variables

H_s and H_a. The main advantage of such a representation of syntactic and semantic rules lies in the simplicity of the recognition algorithm.

4.1. Context-independent algorithm

The following assumptions are made:

1. The name of the variable

$$H_s = \{10 \text{ ms speech frame}\} \tag{2}$$

2. The set of composite terms consists of the names of 5 broad categories corresponding to the stationary parts of 5 chosen broad acoustic articulatory classes and the complementing term '**n**':

$$T_s = \{\text{'r'}, \text{'o'}, \text{'u'}, \text{'s'}, \text{'v'}, \text{'n'}\} \tag{3}$$

3. The universal set U consists of measured and computed parameters

$$U_s = \{U_1 \ldots U_7\} \tag{4}$$

where: $U_1 = A\phi_i$; $U_2 = LP_i$; $U_3 = HP_i$;
$U_4 = A\phi_{\max} - A\phi_i = DA\phi$,
$U_5 = LP_{\max} - LP_i = DLP$,
$U_6 = HP_i - A\phi_i = DHA$,
$U_7 = HP_i - LP_i = DHL$, $i = 1 \ldots N$

N – number of frames, max – denotes maximum values of the corresponding parameter for each word.

The relations defining 5 classes of composite terms and the complementing term are expressed as follows:

DEFINITION 2
The term 'r' describing the acoustic characteristics of the class RS

$$\mathbf{r} = l(DAO_n) \circ [l(DLP_n) \vee l(DHL_n)]. \tag{5}$$

DEFINITION 3
The term 's' describing the acoustic characteristics of the class *SS*

$$\mathbf{s} = h(DHL_n) \wedge h(DHA_n). \tag{6}$$

DEFINITION 4
The term 'u' describing the acoustic characteristics of the class *UC*

$$\mathbf{u} = l(AO_n) \wedge l(LP_n) \wedge l(HP_n). \tag{7}$$

DEFINITION 5
The term 'o' describing the acoustic characteristics of the quasi-stationary part of

the class OC

$$\mathbf{o} = \bar{\mathbf{s}} \wedge \bar{\mathbf{u}} \wedge [m(DAO_n) \odot m(DLP_n)] \tag{8}$$

where $\bar{\mathbf{s}}$ and $\bar{\mathbf{u}}$ have the meaning no **s** and no **u** classes.

DEFINITION 6

The term '**v**' describing the acoustic characteristics of the quasi-stationary part of the class VF

$$\mathbf{v} = m(DLP_n) \wedge h(HP_n) \wedge 0. \tag{9}$$

DEFINITION 7

The term '**n**' complementing in the fuzzy sense the description of the variable H_s

$$\mathbf{n} = \overline{\mathbf{r} \vee \mathbf{s} \vee \mathbf{u} \vee \mathbf{o} \vee \mathbf{v}} \tag{10}$$

where l( ), m( ), h( ) – "low", "medium", "high" functions defined over the measured and calculated parameters, $\odot$ – bounded product, [9], $\wedge$ – intersection.

Equations (4)–(10) are used to compute the numerical values of the possibility that the actually analysed speech frame belongs to the classes terms from the set T_s, representing different manners of articulation. It was assumed, that each frame receives the label of the term with the highest possibility

$$h_i \in H_s \quad h_i \text{ receives the label } k \Leftrightarrow \text{poss}_k > \text{poss}_l \tag{11}$$

where: $k, l \in T_s$,

The membership functions which are used in Eqs. (4)–(10) were estimated heuristically on the set of 120 tokens and then verified on the whole material from the learning group (Fig. 3).

4.2. Context-dependent algorithm

For the context-dependent algorithm, the following assumptions are made:

1. The name of the variable is,

$$H_a = \{\text{"acoustic articulatory segment"}\},$$

2. The set of composite terms which are the labels of 9 broad classes from Fig. 2,

$$T_a = \{'RS', 'RC', 'SS', 'VF', 'OC', 'AR', 'UC', 'PL', 'PS', 'NN'\}$$

3. The universal set U_a

$$U_a = \{'\mathbf{r}', '\mathbf{o}', '\mathbf{u}', '\mathbf{s}', '\mathbf{v}', '\mathbf{n}', U_1\}$$

4. Syntactic rules have the form of the FA network where the input alphabet is U_a and output alphabet is T_a.

5. Semantic rules are the output description functions which assign to each

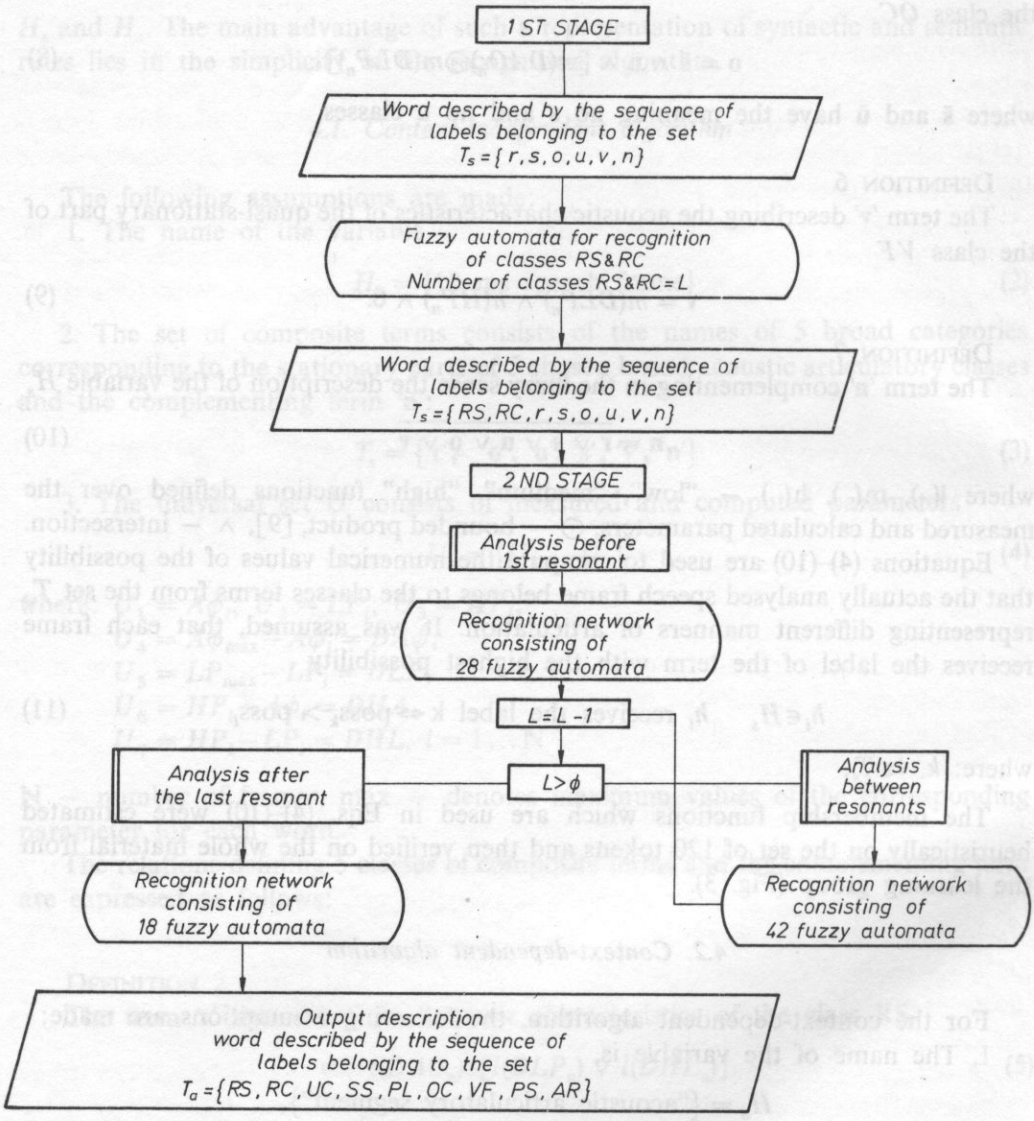

FIG. 3. Two stage context-dependent algorithm for articulatory manner description of speech segments

output label t_a the possibility values stating to what extent the actually analysed acoustic events represented as labels belonging to T_s correspond to the characteristics of 9 broad classes of segments.

The structure of the network is shown in Fig. 3. It is organised as a two-stage network where the whole set of input labels, the result of the previous algorithm, is analysed twice. At the first stage the fuzzy automaton recognizes the classes *RS* and *RC*. Both classes of segments are treated as the strong, "anchor" points of the

analysis (they were detected at 1% error) and around them a more careful analysis of the previously labeled segments is performed to identify classes more complex to detect. At the second stage the algorithm has three separate parts of word analysis viz., before the first resonant in the word, between two consecutive resonants and after the last resonant. The resulting description consists in formulating sequence of labels belonging to the set T_a, with corresponding possibilities of the existence of a class in a given part of the word. At each position of the output string, fuzzy description one or more labels (up to three in the actual network). An example of such a representation of the word /drukarka/ (printer) uttered by a male voice is shown in Fig. 4.

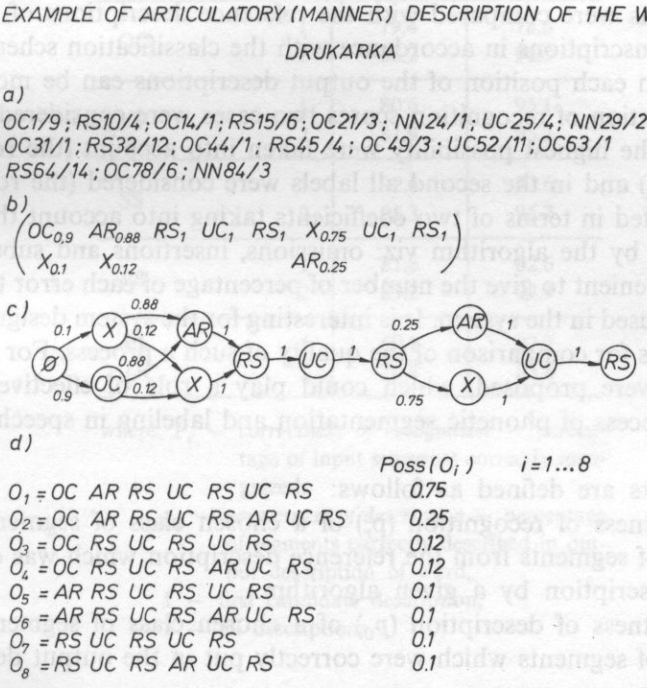

FIG. 4. Example of articulatory manner description of the word /drukarka/ (printer) a) after context-independent analysis, b) after context-dependent analysis, c) graph representation of possible output descriptions, d) chart of output sequences with corresponding degrees of possibility

There are several important advantages of such an approach. The analysis of acoustic events which correspond to the chosen classes of segments is carried independently of the word which is currently analysed, but it depends on their phonetic context. Therefore, the rules of description and, in consequence, the recognition automata are vocabulary-independent. Another important feature of the algorithm is that the same automaton can be used in different contexts so it reduces the amount of memory required to save the state transition tables of automata. The

total number of automata in the network is 52 with 500 states and 1100 paths between them. Finally, the structure of the network is designed in such a way that one can add a new automata in the process of learning when a certain sequence of segments cannot be recognized by the existing network. The actual network was elaborated during the learning process based on a 60 — word vocabulary spoken by 6 persons, containing about 1500 segments.

5. Experimental results

The algorithm was tested on 3120 segments from the 60 word vocabulary spoken by 12 persons (learning + test groups). To verify the efficiency of the algorithm, the output descriptions were compared with the reference descriptions of words made from phonetic transcriptions in accordance with the classification scheme shown in Fig. 2. Because on each position of the output descriptions can be more than one label in the evaluation of recognition scores two cases were considered. In the first only labels with the highest possibility were taken into account (the rows denoted with (1) in Tab. 1) and in the second all labels were considered (the rows (1)). The results are presented in terms of two coefficients taking into account three types of errors committed by the algorithm viz: omissions, insertions and substitutions. It seems very inconvenient to give the number of percentage of each error type for each class of segments used in the system. It is interesting for the system designer just to be apply a good basis for comparison of the quality of such a process. For this purpose two coefficients were proposed, which could play a role of effective parameters describing the process of phonetic segmentation and labeling in speech recognition applications.

The coefficients are defined as follows:

1. The correctness of recognition (p_r) of a chosen class of segments gives the relative number of segments from the reference description which was correctly put in the output description by a given algorithm.

2. The correctness of description (p_0) of a chosen class of segments gives the relative number of segments which were correctly put in the output description by a given algorithm.

The interpretation of either coefficient is the following. The correctness of recognition (p_r) says how difficult the class is for recognition by a given algorithm while the correctness of the description (p_0) shows to what degree one can be sure that the labels representing a given class in the output description belongs in reality to this class of segments.

The results of recognition obtained for 3120 segments (60 words and 12 speakers) are shown in Fig. 5. The average recognition scores were 97% for correctness of recognition and 96% for correctness of description for all the analyzed material. Errors were uniformly spread between resonants and obstruents. For resonants 1 to 5% of segments were badly described, for obstruents this value reached 20 to 25%. Omissions which were the main source of errors were caused by the segments of

Table 1. Results of recognition of broad articulatory classes

		P_r%	P_0%
RS	1	98.3	97.9
	2	99.1	99.3
RC	1	88.8	94.1
	2	99.3	100.0
AR	1	41.7	93.5
	2	56.2	97.5
OC	1	79.4	78.6
	2	88.2	90.0
VF	1	80.6	95.1
	2	81.9	95.2
SS	1	92.6	92.6
	2	95.3	95.2
PS	1	81.3	82.6
	2	89.2	94.3
PL	1	69.1	92.2
UC	2	70.8	95.8

where: P_r — *correctness of recognition* — percentage of input segments correctly recognized,

p_o — *correctness of description* — percentage of segments correctly described in output description of word,

1 — first candidate description,

2 — all descriptions.

classes AR and PL (83% of all omissions). The results are considered to be quite good, given the very simple parametric representation applied in the speech signal.

6. Concluding remarks

Very simple methodological assumptions have led us to the theoretically simple and computationally efficient method of identification of broad classes of segments in IWR. Direct application of the notion of linguistic variable and the process of its description in the bottom-up strategy of word recognition give good results also on the lexical level.

The recognition of the 30–word set was done on 360 utterances spoken by 12

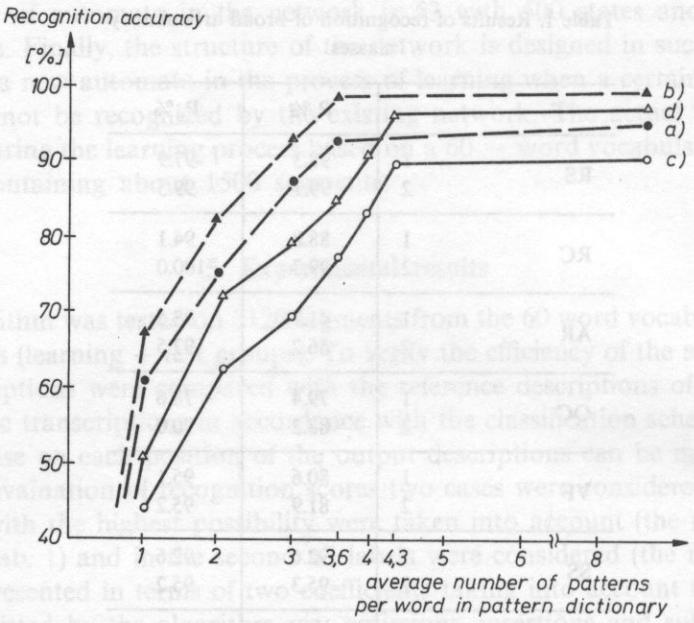

FIG. 5. Results of recognition for 30 words' set a recognition accuracy (R.A.) for the learning group-deterministic approach, b) R.A. for the learning group-fuzzy approach, c) R.A. for test group-deterministic approach, d) R.A. for test group-fuzzy approach

speakers of which 6 formed the training group and 6 speakers were new to the system. When only the highest possibilities were taken into account in the output description case (1), individual recognition scores were between 83–100%; however, when all the descriptions were considered, case (2), they reached 90–100%. Average recognition accuracies were from 91–95% respectively. The 5% increase of the word recognition score is consistent with the results given by De Mori [6]; however, the increase of the recognition rates of certain segment classes is even higher up to 10% (in individual cases to 25%).

To maintain the straightforward bottom-up system architecture, we avoided, in more complex cases, the very useful concept of knowledge source. The subjective and objective knowledge is contained in the FA network which is designed during the learning process. The explicit application of knowledge source will be needed when other levels of speech signal description are incorporated in the system.

An important feature of the method is that the segmentation problem was overcome in such a way that the algorithm estimates the possibility of the presence of a class in a word without explicit description of segment boundaries.

References

[1] W. BEZDEL, *Some problems in man-machine communication using voice*, Int. Journ. Man–Machine Stud. **2**, 257–269 (1970).

[2] M. R. SAMBUR, L. R. RABINER, *A speaker-independent digit recognition system*, Bell Syst. Techn. Journ. **54**, 1, (1975).

[3] C. J. WEINSTEIN, S. S. MC CANDLESS, L. F. MONDSHEIN, Y. W. ZUE, *A system for acoustic-phonetic analysis of continuous speech*, IEEE ASSP **23**, 1, 54–67 (1975).

[4] S. MAKINO, K. KIDO, *Recognition of phonemes using time-spectrum pattern*, Speech Communication **5**, 255–287 (1986).

[5] T. G. von KELLER, *An on-line recognition system for spoken digits*, JASA 49, 1288–1296 (1971).

[6] R. DE MORI, *Computer models of speech using fuzzy algorithms*, Plenum Press, New York, London 1983.

[7] R. GUBRYNOWICZ, *Application de la théorie des sous-ensembles flous à l'analyse et à la reconnaissance automatique de la parole*, Note Technique CNET-LANNION NT/LAA/TSS/157 1983.

[8] W. W. WIĘŹLAK, *Application of approximate articulatory description of speech signal to the recognition of limited set of isolated words* (in Polish) Dr. Eng. Thesis, Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw 1987.

Received August 10, 1989