ARCHIVES OF ACOUSTICS
19, 2, 173 – 183(1994)

THE EFFECT OF FORMANT LEVELS ON THE PERCEPTION OF SYNTHETIC VOWEL SOUNDS

H. KUBZDELA, M. OWSIANNY

Institute of Fundamental Technological Research,
Polish Academy of Sciences,
Department of Acoustic Phonetics
(61-704 Poznań, ul. Noskowskiego 10)

A computer model of a generator of complex periodic sounds simulating vowels has been developed. The system makes it possible to independently control each formant level and to immediately generate the signal. Spectrum envelope within the formant range is approximated by means of a trapezoid contour. A group of 6 listeners used the system to select synthesis parameters for sounds judged by each of them to be maximally similar to the Polish vowels. The sounds were subsequently subjected to evaluation by a more numerous panel of listeners, whereby sounds most suitable for prototypes of Polish vowels were selected. By controlled manipulation of the secont and third formant levels, new sounds were obtained from the prototypes, which, after randomization, were presented for identification to seven listeners. The identification results obtained were presented in tabular form in three variants and characterized from the point of view of the needs of automatic recognition of vowels in continuous speech.

1. Introduction

The phone is assumed in phonetics to be the basic speech segment. In work on automatic speech recognition, such acoustic features of the speech signal have been sought that would define individual phones in a unique way. The knowledge of the role in phone perception of some of their spectral features may be helpful in this search. The significant role of frequencies of some formants in the identification of phone, especially vowels, is commonly recognized. However, somewhat less is known about the significance of formant levels and bandwidths. The effect of formant levels on the perception of vowel sounds can only be examined using a fast synthesizer with independent control of the levels of individual formants. Available synthesizers do not jointly meet both requirements. Therefore, for the purposes of the present work a model of an original vowel synthesizer had been programmed, which was used to study the effect of formant levels on the perception of isolated synthetic vowel sounds.

The paper describes the synthesizer model, the way in which the experiment was run and the obtained results, which will be used in automatic speech recognition.

2. The aim and scope of the study

Not all of distinct spectral features of a phone are indispensable to its correct perceptual identification. Some phone types are characterized by considerable quantitative variation in different spectral features. Studies of the acoustic features of speech sounds, started several tens years ago, were initially concerned with describing and characterizing spectra of those sounds. At that early stage, evaluation of contribution of the various spectral features of phones to their perceptual distinctness was disregarded. The speech synthesizer, capable of imitating speech sounds, opened up new research prospects. The synthesis of individual phones was programmed using data collected in spectrographic analyses of speech. The model of a cascade synthesizer, based on Fant's acoustic theory of speech production, became the standard. A number of fundamental studies in speech synthesis were conducted using those first synthesizers under conditions imposing considerable limitations. A real breakthrough in speech synthesis devices took place due to computerization. Computers opened up new possibilities of using the idea of speech signal synthesis. New expectations addressed to speech synthesis are expressed by researchers preoccupied with automatic recognition of continuous speech. Among others, the synthesizer makes it possible to investigate the role of formant levels in the perceptual identification of vowels. This problem was dealt with in the work reported. The task was to determine the areas of perceptual identification of synthetic Polish vowels, defined by relative levels of the first three formants of optimum mid frequencies. That task was to be carried out using synthetic sounds imitating Polish vowels. For that purpose, a synthesis system was implemented which satisfied two essential conditions. Firstly, $\Delta t1$ and $\Delta t2$ time intervals necessary to supply synthesis parameters and to determine the momentary value of the acoustic signal imitating a vowel should be short enough for the generation of the synthetic sound to take place without a noticeable delay. Therefore, conditions had to be created which would make storing acoustic signal files unnecessary. Since a considerable number of synthetic sounds were to be presented to listeners in random order in a series of perceptual tests, using acoustic signal files instead of synthesis parameters would entail tremendous memory requirements. Secondly, the synthesizer should make it possible to independently control the level of each of the lowest three formants. This was the most essential condition of the study.

3. A stationary vowel synthesizer (SVS)

The two above conditions are not satisfied either by serially produced IC synthesizers based on the cascade model, ruling out independent control of formant levels, or by a software model of Klatt's synthesizer, programmed in a higher level language and, therefore, producing synthetic speech in time significantly exceeding real time. This situation necessitated the development of our synthesis system. It was assumed that for future research a software synthesizer based on the Klatt's model would be developed, implemented at the level of the assembler of the 386 processor. Temporarily, a simplified synthesis model was implemented, generating complex periodic sounds of constant fundamental frequency. A trapezoid approximation of formant contours was adopted in this model. The general principle of this approximation is illustrated in Fig. 1.

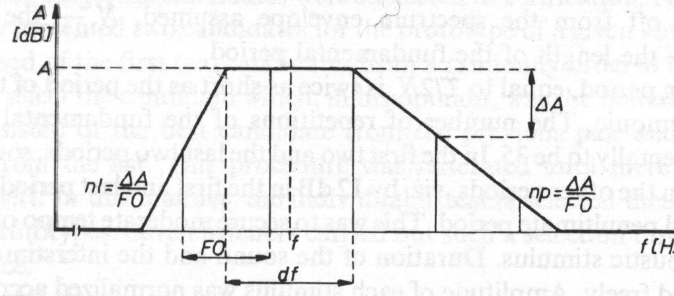


Fig. 1. Illustration of the trapezoid approximation of a formant contour.

The length and position of the upper base of the trapezoid correspond to the width and frequency of the formant. The midpoint of the upper base was assumed to correspond to the centre frequency of the formant. Apart from parameters which refer to the upper base of the trapezoid approximating the spectrum envelope in the formant region, slope parameters are also required. A principle was assumed according to which a spectrum whose envelope is to be approximated by a trapezoid is presented in a linear frequency scale and a decibel intensity scale. Adopting linear frequency scale in this type of approximation should cause those harmonics of the synthetic imitation of an isolated vowel which lie outside the formant regions to be weaker than in the natural vowel. This fact should not essentially affect the results of the experiment testing the significance of formant levels for the perceptual identification of vowels. Figure 2 illustrates the principle of spectrum envelope approximation between adjoining formants.

The first three formants and the contour of the initial segment of the spectrum, covering the lowest harmonics, were approximated in this manner. Thus, the spectrum

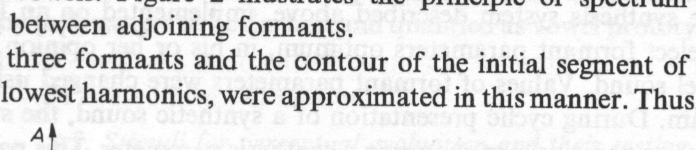


Fig. 2. Illustration of the approximation of a spectrum envelope with two formants.

envelope of a synthetic sound was determined by four trapezoid contours. The synthesis was carried out by summing harmonic function values according to the formula:

$$S_n = \sum_{k=1}^{N} A_k \sin\left(2\pi f k \frac{nT}{2N}\right), \tag{1}$$

where the following symbols denote:

S_n — the n-th sample of the synthetic signal, A_k — the amplitude of the k-th harmonic, read off from the spectrum envelope assumed, N — the number of harmonics, T — the length of the fundamental period.

The sampling period, equal to $T/2N$, is twice as short as the period of the N-th (i.e. the highest) harmonic. The number of repetitions of the fundamental period was selected experimentally to be 35. In the first two and the last two periods, sound intensity was lower than in the other periods, viz. by 12 dB in the first and last period and by 6 dB in the second and penultimate period. This was to secure moderate tempo of the rise and decay of the acoustic stimulus. Duration of the sound and the interstimulus interval could be changed freely. Amplitude of each stimulus was normalized according to the principle that the maximum momentary value was the same for all the stimuli.

4. Assumptions and the course of experiment

It was assumed that perceptual evaluation would be limited to sounds produced from synthetic vowel prototypes and differing from them in the second and third formant levels. The experiment was carried out in a fully automatic way in three stages, each of which had a separate program.

4.1. Prototypes of the isolated synthetic vowels

At the first stage, the prototypes of isolated synthetic vowels were constructed. This task was undertaken by 6 persons with experience in acoustic phonetics. Each of them used the synthesis system described above, implemented on an IBM 386, to individually select formant parameters optimum, in his or her opinion, for a given synthetic vowel sound. Values of formant parameters were changed using a special editing program. During cyclic presentation of a synthetic sound, the subject could effect a positive or negative step change in a synthesis parameter. This parameter had to be selected using the editor. A pair of sounds differing only in this parameter value were repeatedly presented to the listener. The first sound served as the reference for the second (test) sound in which step changes in the parameter indicated could be introduced. If the listener judged the reference sound to be perceptually closer to a given vowel than the test sound he could accept it as a new reference. The parameter with respect to which the two sounds varied took on, in both sounds, an identical value, equal to that of the test sound. By proceeding in this way with all the individual parameters, the listener arrived at a set of values which he judged to be optimum for a given vowel sound. If the reference sound was considered perceptually closer to a given vowel than the test sound (second in the pair), the listener could continue effecting positive or negative changes in the parameter selected or he could abandon verification of this parameter. Following this procedure, 6 listeners defined sets of formant parameter values corresponding to sound judged by the listeners to be the best imitations of Polish vowels. Parameter values selected by individual subjects varied. For each vowel, a number of prototype candidates were obtained.

At the next stage, the candidates were subjected to verification. Nine listeners were individually presented two candidates for the prototype of a given vowel. The first pair was composed of the first two candidates from a randomly ordered list. The listener's task was to select the candidate which, in his opinion, was the better. The next pair of stimuli consisted of the best candidate from the previous pair and the consecutive candidate from the list. This procedure was reiterated until there were no further candidates left. In this manner, the individual listeners selected their own candidates for vowel prototypes. Seven listeners carried out such a selection three times, and tow listeners once.

Table 1. Total numbers of selections by listeners of candidates for vowel prototypes. The candidates selected most often are indicated on the right of the Table

can. vow.	PD	LR	JI	HK	BS	MO
i	1	5	$\boxed{13}$	0	3	1
ɨ	0	$\boxed{12}$	5	2	2	2
e	1	0	$\boxed{10}$	1	7	4
a	0	3	4	5	5	$\boxed{6}$
o	0	$\boxed{9}$	4	2	3	5
u	0	7	$\boxed{8}$	7	1	0

On the basis of the responses obtained, the candidates were ordered with respect to the number of selections and the ultimate list of vowel prototypes was arrived at. This list appears in Table 1. Table 2 presents parameter values of synthetic vowel sounds considered by the listeners to be the best and qualified as vowel prototypes to the main experiment.

4.2. *Stimuli for perceptual evaluation and their testing*

At the next stage, listening tests were conducted which were to provide answers to the main problems considered. Accordingly, the second and third formant levels in synthetic vowel prototypes were modified. A group of seven listeners were aurally presented 294 synthetic sounds generated from the above mentioned prototypes of the six Polish vowels. In each sound coming from the *i*-th prototype, the level of the second formant $P_{F2,i}$ and of the third formant $P_{F3,i}$ took on one of the $2m+1$ values following from the dependences:

Table 2. Parameter values of synthetic vowel sounds qualified by the listeners as prototypes.

vowel	Segment 0					Segment 1					Segment 2					Segment 3				
	f	*df*	*A*	*nl*	*np*	*f*	*df*	*A*	*nl*	*np*	*f*	*df*	*A*	*nl*	*np*	*f*	*df*	*A*	*nl*	*np*
i	125	60	−0.5	9	9	219	60	0	8	8	2600	100	−8.7	5	6	3274	245	−5.4	3	2
ɨ	125	60	−4.4	9	9	323	80	0	9	9	1866	100	−5.7	9	9	2585	100	−5.2	9	5
e	125	60	−6	8	8	580	80	0	8	7	1800	100	−6	7	6	2700	150	−8.8	6	2
a	125	60	−5.1	9	9	850	80	0	9	9	1200	100	−1.6	9	9	2785	200	−6.7	9	3
o	125	60	−3.1	9	9	637	80	0	9	9	987	100	−3.1	9	9	2799	100	−11.7	9	8
u	125	60	−8.9	7	7	300	75	0	9	9	620	90	−9.9	9	8	2400	174	−30.9	4	2

Values are expressed in the following units: *f* [Hz], *df* [Hz], *A* [dB], *nl* [dB/Hz], *np* [dB/Hz].

$$P_{F2,i,j} = P_{F2,i} + j\Delta p\,, \tag{2}$$

$$P_{F3,i,j} = P_{F3,i} + k\Delta p\,, \tag{3}$$

where

$$j \in \{-m, -m+1, \ldots, -1, 0, 1, \ldots m-1, m\}$$

and

$$k \in \{-m, -m+1, \ldots, -1, 0, 1, \ldots, m-1, m\},$$

Δp denotes an increase in level, j and k are positive or negative multiples of an increase in level, $P_{F2,i}$ and $P_{F3,i}$ stand for second and third formant levels in the i-th prototype. Assuming that the size of the sequences to which the multiples j and k belong was identical and equal to $2m+1$, the population of synthetic sounds coming from the prototypes of all the vowels (where $i=6$) included, for $m=3$,

$$N = (2m+1)^2 i = 294 \text{ elements}, \tag{4}$$

Elements j, k and i were selected at random. The listener's task was to state to which vowel the sound heard was definitely similar. A no-decision response was allowed in case of doubt. The responses were fed directly in the computer: either an appropriate vowel letter key was presented or the "X" key, when no decision could be made. The experiment described was run three times. Consequently, each listener supplied $3*294=782$ responses.

5. Results and conclusions

Results were first collected in tables separate for each prototype from which the sound was derived and for each listener. Consequently, 42 tables (7 listeners * 6 vowels) were obtained. One of them is here presented as Table 3. Each column of

Table 3. Results of auditory classification of sounds derived from the prototype of / u /. The responses were obtained from one of the subjects in three listening series. The vowel letters occurring in the Table correspond to appropriate identification responses. y is equivalent to the phonemic symbol ɨ.

Listener: BS
vowel: u

	A2 [dB]	−18	−12	−6	0	+6	+12	+18
A3 [dB]	+18	yyi	yyi	yyy	yyy	uue	ooo	ooo
	+12	iyi	yyi	uyy	uuu	uuu	ooo	ooo
	+6	yyi	yyu	uyu	uuu	uuu	uou	ooo
	0	uuu	uuu	uuu	uuu	uuu	ooo	ooo
	−6	uuu	uuu	uuu	uuu	uuu	ooo	ooo
	−12	uuu	uuu	uuu	uuu	ouu	oou	ooo
	−18	uuu	uuu	uuu	uuu	ouu	ooo	ooo

Table 4. Pooled results of identification of sounds derived from the prototype of / o / by modification of relative levels of the second and third formants. The vowel letters occurring in the Table correspond to appropriate idenfitication responses and the X symbol denotes a no-decision response. Responses obtained from all the listeners are considered. y is equivalent to the phonemic symbol ɨ.

vowel: o

A3 [dB]	−18	−12	−6	0	+6	+12	+18
+18	i: 0 y: 0 e: 13 a: 1 o: 5 u: 0 x: 2	i: 0 y: 0 e: 8 a: 1 o: 10 u: 0 x: 2	i: 0 y: 0 e: 6 a: 3 o: 12 u: 0 x: 0	i: 0 y: 0 e: 5 a: 9 o: 6 u: 0 x: 1	i: 0 y: 0 e: 0 a: 16 o: 5 u: 0 x: 0	i: 0 y: 0 e: 1 a: 19 o: 0 u: 0 x: 1	i: 0 y: 0 e: 1 a: 13 o: 7 u: 0 x: 0
+12	i: 0 y: 0 e: 6 a: 0 o: 14 u: 0 x: 1	i: 0 y: 0 e: 7 a: 0 o: 13 u: 0 x: 1	i: 0 y: 0 e: 4 a: 0 o: 17 u: 0 x: 0	i: 0 y: 0 e: 1 a: 9 o: 11 u: 0 x: 0	i: 0 y: 0 e: 0 a: 17 o: 4 u: 0 x: 0	i: 0 y: 0 e: 0 a: 20 o: 1 u: 0 x: 0	i: 0 y: 0 e: 0 a: 12 o: 8 u: 0 x: 1
+6	i: 0 y: 0 e: 3 a: 0 o: 18 u: 0 x: 0	i: 0 y: 0 e: 2 a: 0 o: 19 u: 0 x: 0	i: 0 y: 0 e: 1 a: 2 o: 18 u: 0 x: 0	i: 0 y: 0 e: 0 a: 10 o: 11 u: 0 x: 0	i: 0 y: 0 e: 0 a: 17 o: 4 u: 0 x: 0	i: 0 y: 0 e: 0 a: 20 o: 1 u: 0 x: 0	i: 0 y: 0 e: 0 a: 14 o: 5 u: 0 x: 2
0	i: 0 y: 0 e: 0 a: 0 o: 21 u: 0 x: 0	i: 0 y: 0 e: 2 a: 1 o: 18 u: 0 x: 0	i: 0 y: 0 e: 1 a: 2 o: 17 u: 0 x: 1	i: 0 y: 0 e: 0 a: 11 o: 10 u: 0 x: 0	i: 0 y: 0 e: 0 a: 18 o: 3 u: 0 x: 0	i: 0 y: 0 e: 1 a: 19 o: 1 u: 0 x: 0	i: 0 y: 0 e: 0 a: 11 o: 10 u: 0 x: 0
−6	i: 0 y: 0 e: 1 a: 0 o: 20 u: 0 x: 0	i: 0 y: 0 e: 1 a: 1 o: 19 u: 0 x: 0	i: 0 y: 0 e: 0 a: 1 o: 20 u: 0 x: 0	i: 0 y: 0 e: 0 a: 14 o: 6 u: 0 x: 1	i: 0 y: 0 e: 0 a: 13 o: 8 u: 0 x: 0	i: 0 y: 0 e: 0 a: 18 o: 3 u: 0 x: 0	i: 0 y: 0 e: 0 a: 15 o: 6 u: 0 x: 0
−12	i: 0 y: 0 e: 0 a: 0 o: 20 u: 0 x: 1	i: 0 y: 0 e: 1 a: 0 o: 20 u: 0 x: 0	i: 0 y: 0 e: 0 a: 3 o: 18 u: 0 x: 0	i: 0 y: 0 e: 0 a: 10 o: 11 u: 0 x: 0	i: 0 y: 0 e: 0 a: 15 o: 6 u: 0 x: 0	i: 0 y: 0 e: 0 a: 19 o: 2 u: 0 x: 0	i: 0 y: 0 e: 0 a: 12 o: 9 u: 0 x: 0
−18	i: 0 y: 0 e: 0 a: 0 o: 21 u: 0 x: 0	i: 0 y: 0 e: 1 a: 0 o: 20 u: 0 x: 0	i: 0 y: 0 e: 0 a: 2 o: 19 u: 0 x: 0	i: 0 y: 0 e: 0 a: 5 o: 16 u: 0 x: 0	i: 0 y: 0 e: 0 a: 15 o: 6 u: 0 x: 0	i: 0 y: 0 e: 0 a: 16 o: 5 u: 0 x: 0	i: 0 y: 0 e: 0 a: 11 o: 10 u: 0 x: 0

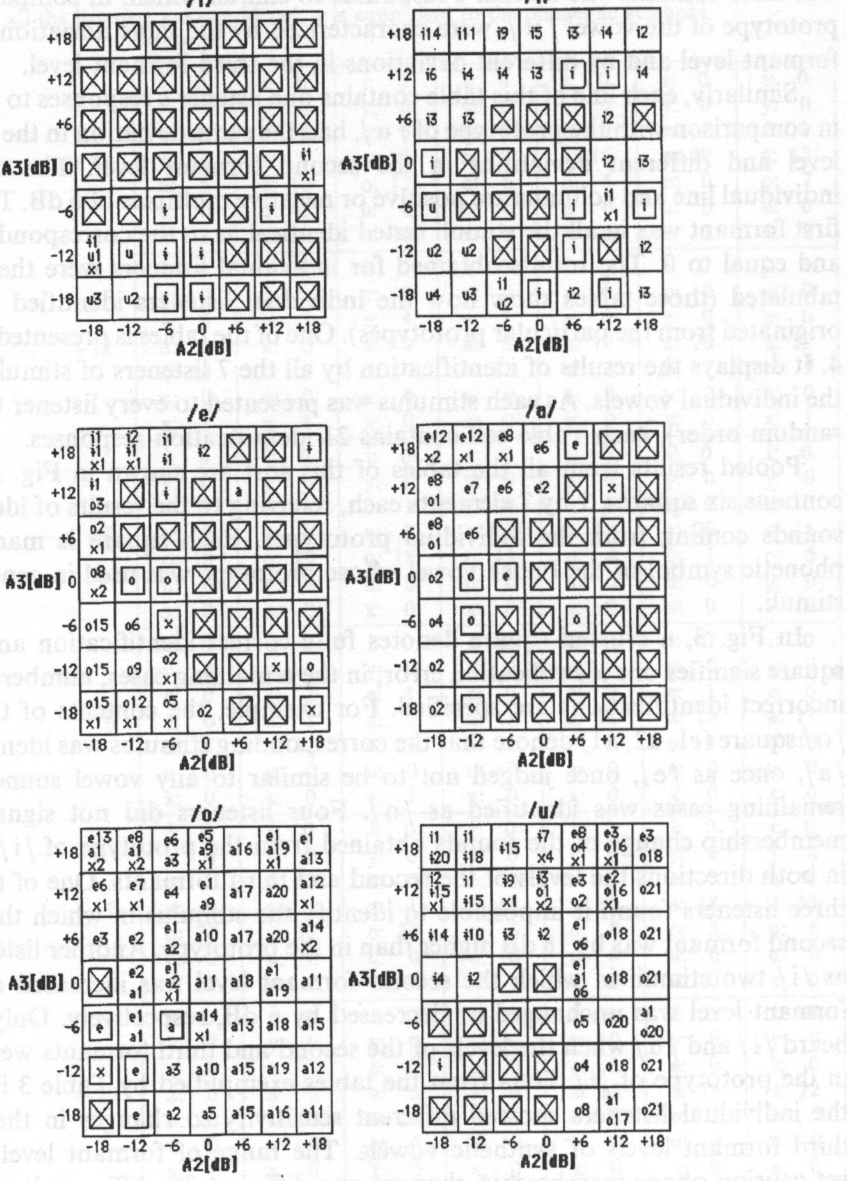
A2 [dB]

this table contains one listener's responses to stimuli which, in comparison with the prototype of the vowel / u /, were characterized by the same deviation in the second formant level and by different deviations in the third formant level.

Similarly, each line of this table contains one listener's responses to stimuli which, in comparison with the prototype of / u /, had the same deviation in the third formant level and different deviations in the second formant level. The deviations for individual line and columns are positive or negative multiples of 6 dB. The level of the first formant was in all the stimuli tested identical as in the corresponding prototype and equal to 0. The results obtained for individual listeners were then pooled and tabulated (those tables show how the individual listeners identified stimuli which originated from the particular prototypes). One of the tables is presented here as Table 4. It displays the results of identification by all the 7 listeners of stimuli representing the individual vowels. As each stimulus was presented to every listener three times (in random order), each Table cell contains 21 identification responses.

Pooled results from all the tables of this sort are shown in Fig. 3. This figure contains six squares, 7 by 7 elements each, referring to the results of identification of sounds coming from the individual prototypes. Each square is marked with the phonetic symbol of the Polish vowel whose prototype was used in generation of the stimuli.

In Fig. 3, a crossed square denotes fully correct identification and an unfilled square signifies one identification error; in the remaining cases, numbers and types of incorrect identifications are specified. For example, the contents of the cell of the / o / square (*e*1, *a*2, *x*1) denote that the corresponding stimulus was identified twice as / a /, once as / e /, once judged not to be similar to any vowel sound, and in the remaining cases was identified as / o /. Four listeners did not signal any phone membership change of the sounds obtained from the prototype of / i / by deviation in both directions the levels of the second and third formants. One of the remaining three listeners found it impossible to identify the stimulus in which the level of the second formant was by 18 dB higher than in the prototype. Another listener classified as / ɨ / two stimuli in which the second formant level was increased and the third formant level was unchanged or decreased by 6 dB, respectively. Only one listener heard / ɨ / and / u / when the levels of the second and third formants were lower than in the prototype of / i /. Data from the tables exemplified by Table 3 indicated that the individual listeners showed different sensitivity to changes in the second and third formant levels of synthetic vowels. The range of formant level changes not yet causing phone membership changes was different for different listeners. For all the listeners, it turned out the widest in the case of / i /, somewhat narrower with / ɨ /, / e / and / a /, and distinctly narrower with / o / and / u /. This last result indicates that in the vowels / o / and / u / the levels of the second and third formant plays an essential role. A surprising result was obtained of identification of the stimulus which, at the previous stage, had been judged by the majority of listeners to be the best candidate for the prototype of / o /.



/i/

A3[dB]	-18	-12	-6	0	+6	+12	+18
+18	☒	☒	☒	☒	☒	☒	☒
+12	☒	☒	☒	☒	☒	☒	☒
+6	☒	☒	☒	☒	☒	☒	☒
0	☒	☒	☒	☒	☒	☒	ɨ1 x1
-6	☒	☒	ɨ	☒	☒	ɨ	☒
-12	ɨ1 u1 x1	u	ɨ	ɨ	☒	☒	☒
-18	u3	u2	ɨ	ɨ	☒	☒	☒

A2[dB]

/ɨ/

A3[dB]	-18	-12	-6	0	+6	+12	+18
+18	i14	i11	i9	i5	i3	i4	i2
+12	i6	i4	i4	i3	i	i	i4
+6	i3	i3	☒	☒	☒	i2	☒
0	i	☒	☒	☒	☒	i2	i3
-6	u	☒	☒	☒	i	i1 x1	i
-12	u2	u2	☒	☒	i	☒	i2
-18	u4	u3	i1 u2	i	i2	i	i3

A2[dB]

/e/

A3[dB]	-18	-12	-6	0	+6	+12	+18
+18	i1 ɨ1 x1	i2 ɨ1 x1	i1 ɨ1	ɨ2	☒	☒	ɨ
+12	ɨ1 o3	☒	ɨ	☒	a	☒	☒
+6	o2 x1	☒	☒	☒	☒	☒	☒
0	o8 x2	o	o	☒	☒	☒	☒
-6	o15	o6	x	☒	☒	☒	☒
-12	o15	o9	o2 x1	☒	☒	x	o
-18	o15 x1	o12 x1	o5 x1	o2	☒	☒	x

A2[dB]

/a/

A3[dB]	-18	-12	-6	0	+6	+12	+18
+18	e12 x2	e12 x1	e8 x1	e6	e	☒	☒
+12	e8 o1	e7	e6	e2	☒	x	☒
+6	e8 o1	e6	☒	☒	☒	☒	☒
0	o2	o	e	☒	☒	☒	☒
-6	o4	o	☒	☒	o	☒	☒
-12	o2	☒	☒	☒	☒	☒	☒
-18	o2	o	☒	☒	☒	☒	☒

A2[dB]

/o/

A3[dB]	-18	-12	-6	0	+6	+12	+18
+18	e13 a1 x2	e8 a1 x2	e6 a3	e5 a9 x1	a16	e1 a19 x1	e1 a13
+12	e6 x1	e7 x1	e4	e1 a9	a17	a20	a12 x1
+6	e3	e2	e1 a2	a10	a17	a20	a14 x2
0	☒	e2 a1	e1 a2 x1	a11	a18	e1 a19	a11
-6	e	e1 a1	a	a14 x1	a13	a18	a15
-12	x	e	a3	a10	a15	a19	a12
-18	☒	e	a2	a5	a15	a16	a11

A2[dB]

/u/

A3[dB]	-18	-12	-6	0	+6	+12	+18
+18	i1 ɨ20	i1 ɨ18	ɨ15	ɨ7 x4	e8 o4 x1	e3 o16 x1	e3 o18
+12	i2 ɨ15 x1	i1 ɨ15	ɨ9 x1	ɨ3 o1 x2	e3 o2	e2 o16 x1	o21
+6	ɨ14	ɨ10	ɨ3	ɨ2	e1 o6	o18	o21
0	ɨ4	ɨ2	☒	☒	e1 a1 o6	o18	o21
-6	☒	☒	☒	☒	o5	o20	a1 o20
-12	ɨ	☒	☒	☒	o4	o18	o21
-18	☒	☒	☒	☒	o8	a1 o17	o21

A2[dB]

Fig. 3. Results from Table 4 shown in a semi-graphic form.

When presented in random order among other sounds, this stimulus was consistently identified as / a / by as many as three listeners, and somewhat less so (in two cases out of three) by one other listener. This fact implies instability of the perceptual boundary between / a / and / o /. The results presented in tabular – graphic form in Fig. 3 demonstrate to what extent and in what direction changes the perceptual classification of sounds differing from vowel prototypes in the second and third formant levels. Thus, e.g. the prototype of / u / with the second formant level decreased and the third formant level increased was identified as / I /, whereas with the second formant level increased was predominantly perceived as / o /, irrespective of the third formant level. Relations of a similar kind between the result of perceptual identification of a synthetic sound derived from one of the six vowel prototypes and the difference in the second and third formant levels between the sound tested and the prototype also occur when the source from which the sound tested is derived are prototypes of other vowels. Nevertheless, those relations are most conspicuous in the case of the prototypes of / u / and / o /. The results obtained indicate that formant frequencies alone are insufficient for the classification of vowel segments to be recognized. Two vowel sounds with the same formant frequencies can belong to different classes depending on the level of the second and third formant, relative to the first formant level. The conclusions arrived at in the present paper will help to supplement the set of vowel sound features useful in automatic speech recognition. Heretofore, mainly frequency parameters were made use of and the features relating to formant levels were not paid sufficient attention.

References

[1] L.A. CHISTOVICH et al., *Centers of gravity and spectral peaks as the determinants of vowel quality*, Frontiers of Speech Communications Research (ed. B. Lindblom, S. Ohman), Academic Press, 145–157, 1978.

[2] G. FANT, *Acoustic theory of speech production*, Mouton, The Hague, 1960.

[3] G. FANT, et al., *Two formant models, pitch and vowel perception*, Auditory Analysis and Perception of Speech (ed. G. Fant, M.A.A. Tatham), Academic Press, 1975.

[4] T. HOUTGAST, *Auditory analysis of vowel-like sounds*, Acoustica, 4, 320–324 (1974).

[5] D. KLATT, *Representation of the first formant in speech recognition and in models of the auditory periphery*, Proc. of the Montreal Symp. on Speech Recognition, 1986.

[6] K. KUDELA-DOBROGOWSKA, *Further studies of the optimal formant frequency values of Polish vowels*, in: Speech Analysis and Synthesis, vol. 3, (ed. W. Jassem), 265–285, 1973.

[7] V.G. LEBEDEV and N.G. ZAGORUIKO, *Auditory perception and speech recognition*, Speech Communication, 4, 97–103 (1985).

Received April 14, 1993 revised english version November 11, 1993

form in Fig. 3 demonstrate to what extent and in what direction changes the perceptual classification of sounds differing from vowel prototypes in the second and third formant levels. Thus, e.g. the prototype of /u/ with the second formant level decreased and the third formant level increased was identified as /i/, whereas with the second formant level increased was predominantly perceived as /o/, irrespective of the third formant level. Relations of a similar kind between the result of perceptual identification of a synthesized sound derived from one of the six vowel prototypes and the difference in the second and third formant levels between the sound tested and the prototype also occur when the source from which the sound tested is derived are prototypes of other vowels. Nevertheless, these relations are most conspicuous in the case of the prototypes of /a/ and /o/. The results obtained indicate that formant frequencies alone are insufficient for the classification of vowel segments to be recognized. Two vowel sounds with the same formant frequencies can belong to different classes depending on the level of the second and third formant, relative to the first formant level. The conclusions arrived at in the present paper will help to supplement the set of vowel sound features useful in automatic speech recognition. Heretofore, mainly frequency parameters were made use of and the features relating to formant levels were not paid sufficient attention.

References

[1] L.A. Chistovich et al., *Centres of gravity and spectral peaks as the determinants of vowel quality*, Frontiers of Speech Communications Research (ed. B. Lindblom, S. Öhman), Academic Press, 143–157, 1979.

[2] G. Fant, *Acoustic theory of speech production*, Mouton, The Hague, 1960.

[3] G. Fant et al., *Two-formant models, pitch and vowel perception*, Auditory Analysis and Perception of Speech (ed. G. Fant, M.A.A. Tatham), Academic Press, 1975.

[4] T. Houtgast, *Auditory analysis of vowel-like sounds*, Acustica, 4, 320–324 (1974).

[5] D. Klatt, *Representation of the first formant in speech recognition and in models of the auditory periphery*, Proc. of the Montreal Symposium on Speech Recognition, 1986.

[6] K. Kudela-Dobrogowska, *Further studies of the optimum formant frequency values of Polish vowels*, in: Speech Analysis and Synthesis vol. 4, (ed. W. Jassem), 243–258, 1976.

[7] V.G. Lebedev and N.G. Zagoruyko, *Auditory perception and speech recognition*, Speech Communication, 4, 97–103 (1985).

Received April 14, 1987; revised english version September 22, 1988