ARCHIVES OF ACOUSTICS
16, 2, 305-323 (1991)

DETERMINATION OF PERCEPTUAL BOUNDARIES BETWEEN THE MALE FEMALE AND CHILD'S VOICES IN ISOLATED SYNTHETIC POLISH VOWELS[1]

J. IMIOŁCZYK

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

The problem of perceptual boundaries between the male, female and child's voices was considered. The experimental material included 730 synthetic realisations of the six Polish oral vowels: /i/, /ɨ/, /e/, /a/, /o/ and /u/. „Target" male, female and child's utterances as well as a number of intermediate ones were obtained for each vowel by selecting appropriate combinations of FO and formant frequency values. Results of the listening test show that FO is the principal factor determining the perception of voice category and that, as such, it plays the key role in the perceptual normalisation of the speaker's vocal tract.

W pracy podjęto problem granic percepcyjnych między głosami męskim, kobiecym i dziecięcym. Jako materiał badawczy wykorzystano 730 syntetycznych wypowiedzi sześciu ustnych samogłosek polskich: /i/, /ɨ/, /e/, /a/, /o/ oraz /u/. Poprzez odpowiedni dobór wartości FO i częstotliwości formantowych zsyntetyzowano dla każdej z samogłosek „docelowe" realizacje męskie, kobiece i dziecięce oraz szereg realizacji pośrednich. Rezultaty badań odsłuchowych wskazują, że częstotliwość podstawowa jest zasadniczym czynnikiem decydującm o postrzeganej kategorii głosu i że odgrywa ona w związku z tym kluczową rolę w percepcyjnej normalizacji toru głosowego osoby mówiącej.

1. Introduction

Despite many years of continuous research and hundreds of experiments it has not yet been explained in what way variability of the speech signal, posing so great a problem e.g. in automatic speech recognition, is eliminated with remarkable ease and efficiency in the process of perception. In attempts to resolve this issue some authors (notably K. N. Stevens and S. Blumstein) have claimed that in the apparently variable signal there occur certain invariant features which provide guidelines to perceptual classification processes. Other investigators have considered this line of research unproductive and have suggested concentrating on attempts to find out **how** the listener deals with the variability which is there [11].

[1] This research was carried out within the CPBP 02.03 Problem.

One of the sources of variability facing the listener in speech perception are interspeaker differences. As noted by Holmes [20], variations in the acoustic structure of a vowel spoken by a man, woman and child are so considerable that the phonetic equivalence of the three utternances **must** have a perceptual basis. He goes on to say that there probably exist **systematic relations** between acoustic features, which allow interpretation of markedly different acoustic signals as linguistically equivalent (p. 347). Thus, it seems valid to assume that (1) in order to efficiently eliminate the speaker-related variability from the speech signal the listener must be able to differentiate at least three major voice categories, viz. the male, female and child's voice and (2) the voice category information is contained in the signal in the form of appropriate relations between acoustic features.

The possibility of perceptual identification of a voice as male, female or child's implies the existence of more or less sharply defined boundaries between the three voice types. As can be expected, determining those boundaries and establishing which acoustic parameters affect them should contribute to a better understanding of the mechanism of perceptual normalisation of the speaker's vocal tract. These two aims have provided motivation for the present research.

2. Recording, analysis and resynthesis of natural vowels

At the initial stage, six Polish oral vowels, i.e. [i], [ɨ], [e], [a], [o] and [u], spoken by a male voice, were recorded under laboratory conditions. An RS 249–946 microphone and a cassette recorder Revox B710 were used for that purpose. Special care was taken to ensure that the level of the recording and the FO pattern (rising-falling) were the same for all the utterances.

The vowels were then low-pass filtered, sampled at 10 kHz and stored on disk of a MASSCOMP MC5400 computer. In their subsequent acoustic analysis, a specialised software packet, named AUDLAB, was used. Of the functions it included the following were applied:

- FO extraction
- calculation of momentary and average spectra
- preparation of 2- and 3-dimensional spectrograms
- measurement of segment duration.

On the basis of the data derived from the analysis, the vowels were (re)synthesized, using the Klatt software formant snythesiser[2]. Choosing the parallel branch of the synthesiser made it possible to control the amplitudes of individual formants. In order to make the vowels sound as natural as possible, the second of the two voicing sources („ss" = 2) was selected. The rate of D/A conversion was 10 000 samples/sec.

Apart from the „standard" parameters of formant synthesis (eg. FO, bandwidths of spectrum envelope peaks etc.), two additional ones were used. One of them, "no"

[1] A later version of the programme described in [22] was used; see also [1].

corresponds to the number of samples in the open phase of the voicing source period and the other one, "tl", controls the spectral tilt. The principles of handling these two parameters will be described in Sections 3.2 and 4.2.

Vowels synthesized on the basis of the data from the analysis of natural vowels were analyzed again and the spectral characteristics of the two groups of vowels were compared. Necessary corrections in parameter values were introduced until the spectra of synthetic vowels closely matched those of their natural counterparts.

In order to minimize quality-unrelated differences between the particular synthetic vowels, their duration was made uniform (equal to 440 ms), and the same amplitude and FO pattern, extracted from the natural [e][3], were applied in all of them. The values of "no" and "tl" were also made identical.

Consequently, the differences between the synthetic vowels were limited to: (1) frequencies, (2) amplitudes and (3) bandwidths of formants as well as (4) the overall gain control which was used to make the vowels approximately equally loud.

In each vowel, the above-named parameters had constant values, typical of that vowel. Time-varying parameters such as amplitude of the voice source and FO were updated every 5 ms[4].

The FO contour is illustrated in Table 1. Changes between the discontinuity points ($t = 0$ ms, $t = 100$ ms etc.) are linear.

Formant frequencies of the six synthetic vowels are shown in Table 2.

Table 1. FO contour in synthetic male vowels

t [ms]	0	100	245	390	435
FO [Hz]	139.0	144.0	119.0	112.0	119.5

Table 2. Formant frequencies of synthetic male vowels

Formant	Vowel					
[Hz]	i	ɨ	e	a	o	u
F1	250	365	600	800	620	340
F2	2170	1960	1720	1220	950	760
F3	3100	2530	2620	2760	2730	2250
F4	3770	3230	3050	3760	3920	3030
F5	4500	4100	4350	4850	4560	4140

[3] Strictly speaking, both were aproximations of the patterns characterizing the natural [e].

[4] Strictly speaking, those changes were delayed until the beginning of the next glottal pulse (cf. [22], p. 978).

3. Synthesis of female vowels

3.1. *Formant frequencies*

The vowel formant frequencies given in Table 2, assumed to be representative of a male voice, were subsequently scaled in such a way that for each original (male) vowel its phonetic equivalent was obtained with formant frequencies typical of a female voice. The *k* scaling factors given by FANT ([14], p. 87) were used for that purpose.

The length of the supralaryngeal vocal tract is known to be approx. 15...20 per cent shorter in the female than in the male (eg. [5], [14]). However, owing to a difference in proportion of the length of the oral to the pharyngeal cavity in both genders, the frequencies of the male and female formants cannot be related by means of a single factor. As shown by FANT ([14], [15]), in order to express the gender-related differences in the frequencies of the three lower vowel formants, three separate scaling factors (*k1*, *k2* and *k3*) are required which, in addition, depend on vowel type.

Table 3 presents the values of scaling factors adopted in the present work. They are the result of a modification of FANT's data ([14], Table. 1), taking into account the articulatory-acoustic specificity of the Polish vowels. The values of *k4* and *k5*, not considered in [14], were in all cases equal to 17 per cent. Table 3 also shows the female formant frequencies obtained by appropriately increasing the male formants.

Table 3. *k* scaling factors and formant frequencies of synthetic female vowels

Vowel	*k*1 [%]	F1 [Hz]	*k*2 [%]	F2 [Hz]	*k*3 [%]	F3 [Hz]	*k*4 [%]	F4 [Hz]	*k*5 [%]	F5 [Hz]
i	8	270	22	2647	16	3596	17	4411	17	5265
ɨ	10	402	24	2430	20	3036	17	3779	17	4797
e	24	744	20	2064	19	3118	17	3569	17	5090
a	16	928	16	1415	16	3202	17	4399	17	5675
o	15	713	14	1083	15	3140	17	4586	17	5335
u	10	374	5	798	19	2678	17	3545	17	4844

Owing to the differences in the *k* factors for the individual vowels, selecting the formant sets intermediate between the male and the female formant values required that each vowel should be treated separately. It was assumed that the differences in F1, F2 and F3 between the consecutive formant sets, corresponding to a shortening of the vocal tract, would not be greater than the difference limens given by FLANAGAN [16], equal to: ±20 Hz for F1, ±50 Hz for F2 and ±75 Hz for F3. Consequently, the following numbers of formant sets were obtained for each vowel: [i] – 11, [ɨ] – 11, [e] – 9, [a] – 8, [o] – 7 and [u] – 7. The sets are presented in Table 4.

Table 4. Male/female formant frequency sets

Vowel	Formant set	F1, ΔF1 [Hz]	F2, ΔF2 [Hz]	F3, ΔF3 [Hz]	F4, ΔF4 [Hz]	F5, ΔF5 [Hz]
i	1	250	2170	3100	3770	4500
	…	Δ = 2	Δ = 47.7	Δ = 49.6	Δ = 64.1	Δ = 76.5
	11	270	2647	3596	4411	5265
ɨ	1	365	1960	2530	3230	4100
	…	Δ = 3.7	Δ = 47.0	Δ = 50.6	Δ = 54.9	Δ = 69.7
	11	402	2430	3036	3779	4797
e	1	600	1720	2620	3050	4350
	…	Δ = 18.0	Δ = 43.0	Δ = 62.2	Δ = 64.9	Δ = 92.5
	9	744	2064	3118	3569	5090
a	1	800	1220	2760	3760	4850
	…	Δ = 18.3	Δ = 27.9	Δ = 63.1	Δ = 91.3	Δ = 117.9
	8	928	1415	3202	4399	5675
o	1	620	950	2730	3920	4560
	…	Δ = 15.5	Δ = 22.2	Δ = 68.3	Δ = 111.0	Δ = 129.2
	7	713	1083	3140	4586	5335
u	1	340	760	2250	3030	4140
	…	Δ = 5.7	Δ = 6.3	Δ = 71.3	Δ = 85.8	Δ = 117.3
	7	374	798	2678	3545	4844

In the synthesis of all the utterances representing the same vowel, formant amplitudes and bandwidths were constant and identical with those in the original ("male") set. Formant frequencies higher than 5 kHz were not considered.

3.2. *Fundamental frequency*

The fundamental frequency of a female voice is, on average, an octave higher than that a male voice (e.g. [13], p. 242) and amounts to approx. 220 Hz. In the present work, fundamental frequency equal to 1.7 that of the male was assumed, after [23], as typical of a female voice. The female FO pattern and 6 intermediate ones were obtained by multiplying FO values at each of the discontinuity points of the male pattern by an appropriate scaling factor (1.1, 1.2… 1.7). The average FO value was 128 Hz in the male pattern and 218 Hz in the female one (see Table 5).

As shown by a number of authors (eg. [19], [23], [28]), the proportion of duration of the open and the closed phase of the glottal period (the open qotient) is greater in the female voice and the shape of the glottal pulse is more symmetrical. As a consequence, the female voice is characterized by a steeper spectral tilt: whilst the rate of the spectral fall-off in a typical male voice averages −12 dB/octave ([12], [13], [17]), in a female voice, it can be as sharp as −15… −18 dB/octave ([23],

Table 5. Male/female FO contours and glottal period characteristics

scal. fact.	FO [Hz] Oms	100ms	245ms	390ms	435ms	mean FO [Hz]	no	% peri. dur.	tl
1.0	139.0	144.0	119.0	112.0	119.5	128	33	42%	0
1.1	152.9	158.4	130.9	123.2	131.5	141	33	47%	3
1.2	166.8	172.8	142.8	134.4	143.4	154	33	51%	6
1.3	180.7	187.2	154.7	145.6	155.4	166	33	55%	9
1.4	194.6	201.6	166.6	156.8	167.3	179	33	59%	12
1.5	208.5	216.0	178.5	168.0	179.3	192	33	63%	15
1.6	222.4	230.4	190.4	179.2	191.2	205	33	68%	18
1.7	236.6	244.8	202.3	190.4	203.2	218	33	72%	21

[28]). It has to be noted, however, that in the case of phonation at high pitch, the spectral tilt in a male voice can also approach −18 dB/octave ([28]).

In view of the above observations, the following two principles relating to the voice source characteristics were adopted in the present work:

(1) with an increase in mean FO, the spectral tilt became steeper; it was −12 dB/octave for “tl” = 0, and ca. −17 dB/octave for “tl” = 21.

(2) duration of the open phase of the period was in all the cases equal to 3.3 ms (“no” = 33), which means that the open quotient ranged from 42 to 72 per cent.

Each of the eight FO patterns (Table 5) was combined with each of the formant sets distinguished for the individual vowels (Tables 4…9). A total of 424 stimuli were synthesized at this stage.

4. Synthesis of child’s vowels

4.1. *Formant frequencies*

The synthesis of child’s vowels was based on FANT’s claim (cf. [14]) that the differences in formant frequencies between female and child’s vowels can be expressed by means of a single scaling factor independent of vowel type. All the female formant frequencies (i.e. the ones occuring in the last sets in Table 4) were consequently increased by the factor of $k = 18$ per cent (see Table 6).

Table 6. Child’s formant frequencies

Vowel	F1 [Hz]	F2 [Hz]	F3 [Hz]	F4 [Hz]
i	319	3123	4243	5205
ɨ	474	2867	3582	4459
e	878	2436	3679	4211
a	1095	1670	3778	5191
o	841	1278	3705	5411
u	441	942	3160	4183

Apart from the "target" child's vowels, vowels with intermediate (between female and child's) formant frequencies were also synthesized. Similarly as before (cf. section 3.1), it was assumed that $\Delta k1$ will be less than 20 Hz, $\Delta k2 < 50$ Hz, $\Delta k3 < 75$ Hz. As a result, the following numbers of formant sets were obtained for each of the vowels: [i] — 10, [ɨ] — 9, [e] — 8, [a] — 9, [o] — 8 and [u] — 7. The sets are presented in Table 7. Their numbering continues the numbering from Table 4.

Table 7. Female/child's formant frequency sets

Vowel	Formant set	F1, ΔF1 [Hz]	F2, ΔF2 [Hz]	F3, ΔF3 [Hz]	F4, ΔF4 [Hz]
i	12	275	2695	3661	4490
	…	$\Delta = 4.9$	$\Delta = 47.6$	$\Delta = 64.7$	$\Delta = 79.4$
	21	319	3123	4243	5205
ɨ	12	410	2479	3097	3855
	…	$\Delta = 8.0$	$\Delta = 48.5$	$\Delta = 60.6$	$\Delta = 75.5$
	20	474	2867	3582	4459
e	10	761	2111	3188	3649
	…	$\Delta = 16.7$	$\Delta = 46.4$	$\Delta = 70.1$	$\Delta = 80.3$
	17	878	2436	3679	4211
a	9	947	1443	3266	4487
	…	$\Delta = 18.5$	$\Delta = 28.4$	$\Delta = 64.0$	$\Delta = 88.0$
	17	1095	1670	3778	5191
o	8	729	1107	3211	4689
	…	$\Delta = 16.0$	$\Delta = 24.4$	$\Delta = 70.6$	$\Delta = 103.1$
	15	841	1278	3705	5411
u	8	384	819	2747	3636
	…	$\Delta = 9.5$	$\Delta = 20.5$	$\Delta = 68.8$	$\Delta = 91.2$
	14	441	942	3160	4183

As in the previous stage, formant amplitudes and bandwidths were not changed. Formant frequencies exceeding 5 kHz were disregarded.

4.2. *Fundamental frequency*

According to Fant ([13], p. 242), voice fundamental frequency in a child at the age of 10 averages 300 Hz, although the interspeaker variation may be considerable. Similar, though somewhat lower values are quoted by Hasek and Singh [18] for 5…10 year old boys and girls.

Considering the above observations, furhter five FO patterns were synthesized in the way described in section 3.2. The scaling factors used ranged from 1.8 to 2.2. For want of data concerning the glottal tone characteristics of a child's voice, the value of

Table 8. Female/child's FO contours and glottal period characteristics

scal. fact.	FO [Hz] Oms	100ms	245ms	390ms	435ms	mean FO [Hz]	no	% peri. dur.	tl
1.8	250.2	259.2	214.2	201.6	215.1	230	26	60%	21
1.9	264.1	273.6	226.1	212.8	227.1	243	26	63%	21
2.0	278.0	288.0	238.0	224.0	239.0	256	26	67%	21
2.1	291.9	302.4	249.9	235.2	251.0	269	26	70%	21
2.2	305.8	316.8	261.8	246.4	262.9	282	26	72%	21

"tl" was in all cases the same as in the "target" female pattern (cf. Table 5) and the open quotient varied from 80% to 72%. The relevant figures are given in Table 8.

Each of the five FO patterns and, additionally, the pattern with the scaling factor of 1.7 (see Table 5) were combined with each of the formant sets presented in Table 7. As a result, 306 stimuli were obtained.

5. Listening test

5.1. *Test material. The manner of presentation*

The total number of stimuli generated amounted to 730 (424+306). The set comprized: 148 [i]s, 142 [ɨ]s, 120 [e]s, 118 [a]s, 104 [o]s, and 98 [u]s. The material was randomized using a special procedure included in the software package used for the synthesis and recorded on a Revox B710. The inter-stimulus interval (ISI) on the test tape was 3 secs, which, according to [7], is the decay time of the auditory memory in a vowel discrimination task and the time after which context effects disappear in vowel identification.

The interval between groups of 10 stimuli was 4.5 secs.

20 subjects with no known hearing impairments participated in the listening experiment. Their task consisted in identifying each stimulus as one of the six vowels and classifying it as an utterance by a man, a woman or a child. This was done by putting in appropriate two-letter symbols (e.g. "im" – [i], male voice, "af" – [a], female voice, "uc" – [u], child's voice) in an answer sheet.

The material was presented to the subjects in two sessions a few days apart. At the beginning of each session the subjects were instructed as to their task and the set of 18 "target" male, female and child's vowels was played to them in random order.

5.2. *Results*

5.2.1. Vowel identification. Of the total of 14 600 responses 557 were incorrect (3.8%), [o] being by far the most frequently misidentified vowel. Error rates for the individual vowels are given in Table 9.

Table 9. Vowel identification errors

Vowel present.	No of stimuli	No of errors	% errors	misidentified* as
u	1960	2	0.1	−(2)
e	2400	4	0.2	−(4)
i	2960	18	0.6	ɨ (14), u(1), −(3)
a	2360	28	1.2	o(24), −(4)
ɨ	2840	190	6.7	i(145), e(38), −(7)
o	2080	315	15.1	a(312), −(3)

* "−" in this column denotes the lack of response in an answer sheet

Percentages of identification errors for [u], [e], [i] and [a] can be said to comply with the "norm". This is confirmed by the random distribution of incorrect responses obtained for these vowels. On the other hand, a considerable number of mistakes which occurred in the identification of [ɨ] and, especially, [o] seem to point to a "deficiency" in the acoustic structure of (some) stimuli representing these vowels. A probable explanation of this fact will be offered in section 6.1.

5.2.2. Recognition of voice category. When summing up the number of occurrences of voice qualifiers ("m" for male, "f" for female and "c" for child's) ascribed to the particular stimuli, all the responses containing a vowel identification error were ignored. With respect to some stimuli, especially of the [o] type, this considerably limited the number of "effective" voice category recognitions.

Figures 1...6 present the boundaries between the three voice categories deter-

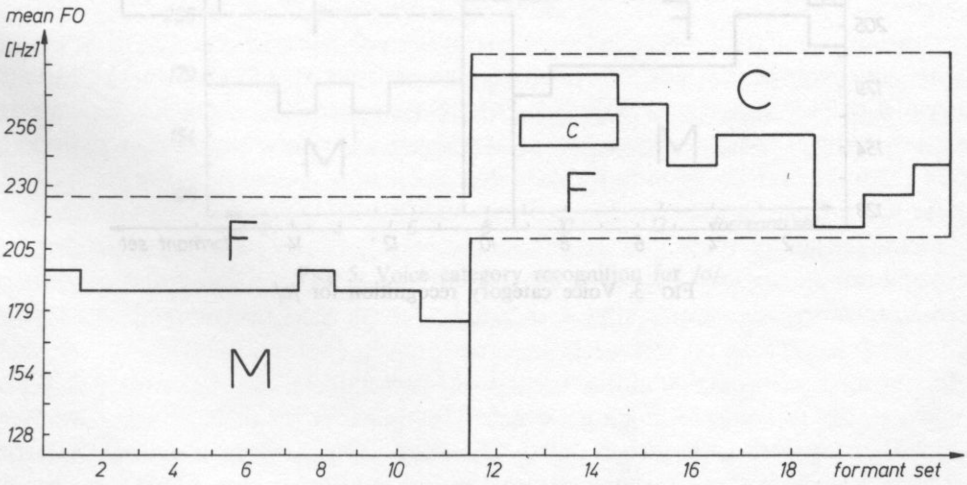

FIG. 1. Voice category recognition for /i/

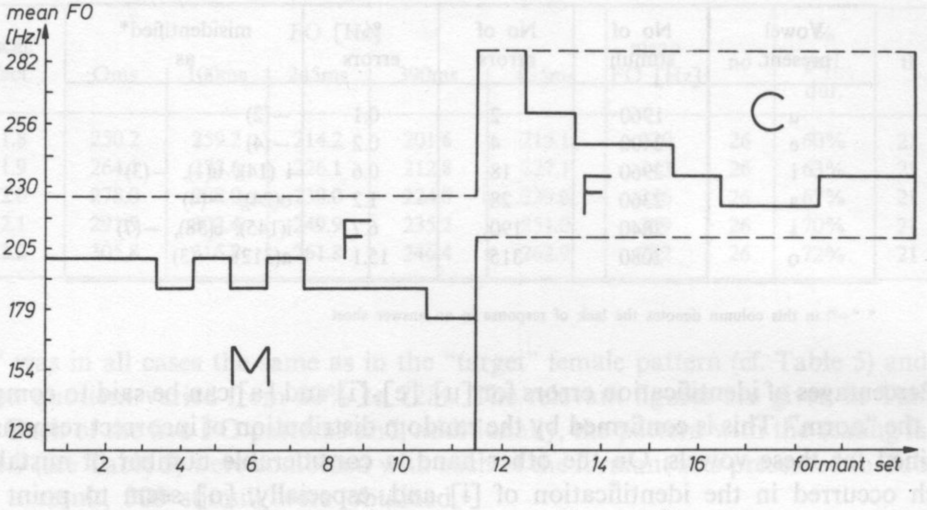

FIG. 2. Voice category recognition for /ɨ/

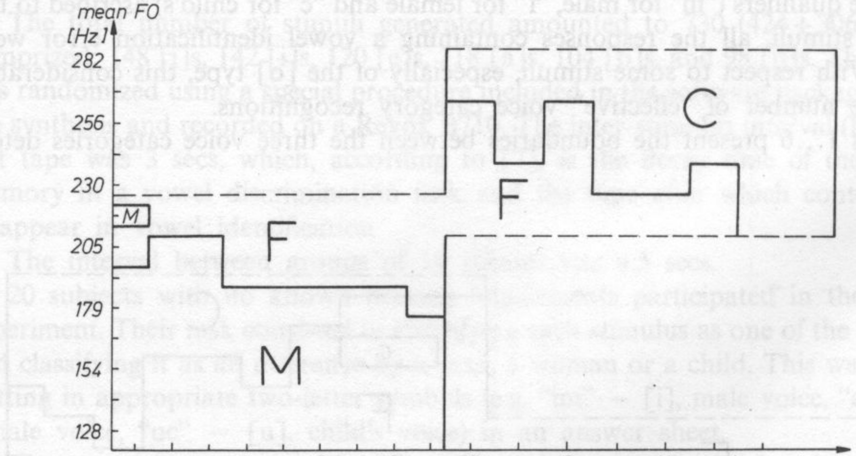

FIG 3. Voice category recognition for /e/

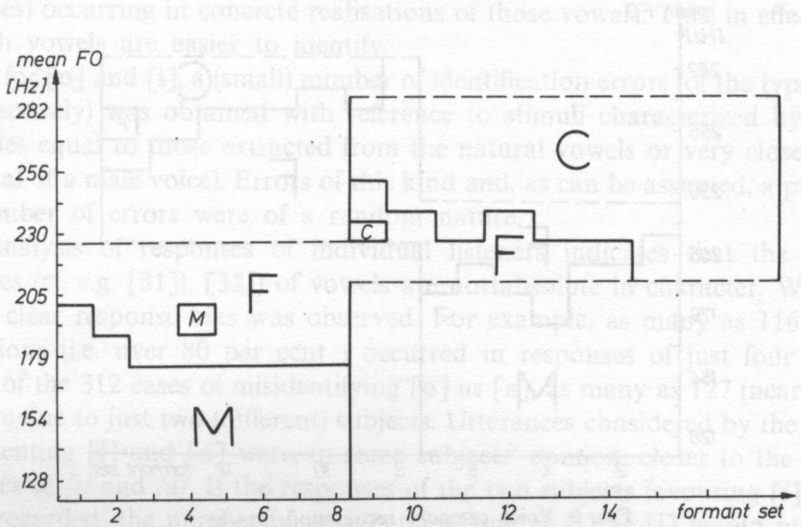

FIG. 4. Voice category recognition for /a/

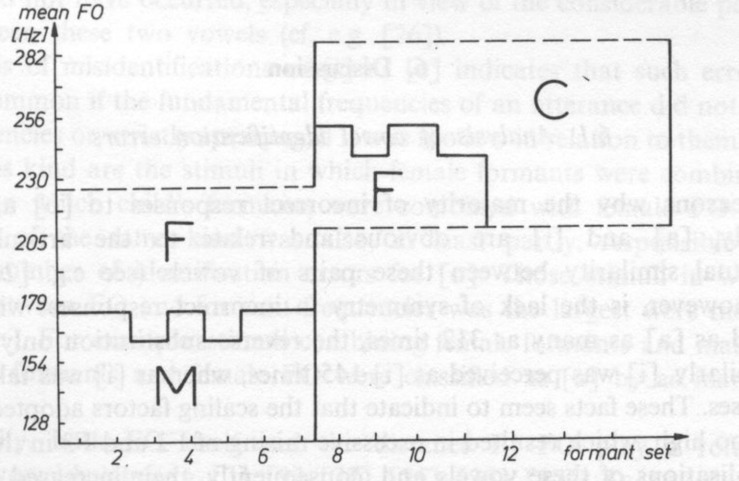

FIG. 5. Voice category recognition for /o/

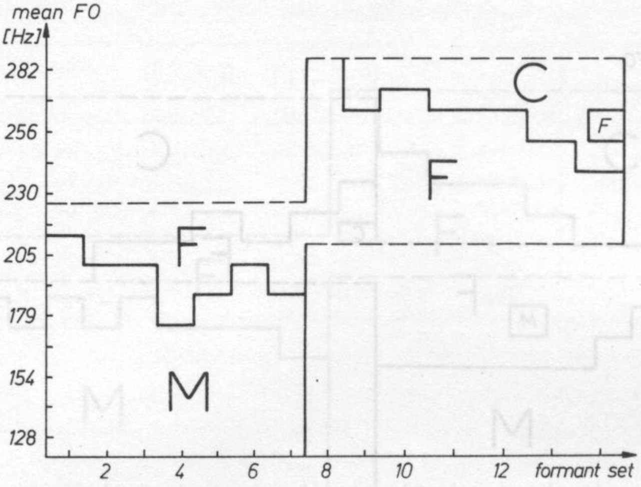

FIG. 6. Voice category recognition for /u/

mined for the individual vowels. The criterion used was the number of "m", "f" or "c" responses dominant for any given combination of vowel formants and an FO pattern.

6. Discussion

6.1. *Analysis of vowel identification errors*

The reasons why the majority of incorrect responses to [o] and [ɨ] were, respectively, [a] and [i] are obvious and relate to the articulatory-acoustic-perceptual similarity between these pairs of vowels (see e.g. [26]). What is striking, however, is the lack of symmetry in incorrect responses: whilst [o] was recognized as [a] as many as 312 times, the reverse substitution only occurred 24 times; similarly, [ɨ] was perceived as [i] 145 times, whereas [i] was taken for [ɨ] in only 14 cases. These facts seem to indicate that the scaling factors adopted for [o] and [ɨ] were too high, which resulted in excessive raising of F2 and F3 in the female and child's realisations of these vowels and, consequently, their increased acoustic and perceptual similarity to [a] and [i]. While such an explanation cannot be categorically dismissed (especially in the case of [o]), a few other facts should be noted which are of some relevance in this context.

According to Stevens' quantal theory of speech [33], [i], [a] and [u] have a special status in spoken language which is manifested by their forming discrete perceptual categories rather than being identified as points on a continuum. As a result, with such (quantal) vowels the perceptual classificatory mechanisms appear to be more tolerant to various acoustic deviations (from the appropriate phonetic prototypes) occurring in concrete realisations of those vowels. This, in effect, means that such vowels are easier to identify.

Both for [o] and [ɨ], a (small) number of identification errors (of the type [a] and [i], respectively) was obtained with reference to stimuli characterized by formant frequencies equal to those extracted from the natural vowels or very close to them (i.e. typical of a male voice). Errors of this kind and, as can be assumed, a part of the total number of errors were of a random nature.

An analysis of responses of individual listeners indicates that the phonetic prototypes (cf. e.g. [31]), [32]) of vowels are not absolute in character. With some subjects, clear response bias was observed. For example, as many as 116 [ɨ] → [i] substitutions (i.e. over 80 per cent) occurred in responses of just four persons; likewise, of the 312 cases of misidentifying [o] as [a], as many as 127 (nearly 42 per cent) were due to just two (different) subjects. Utterances considered by the majority as representing [ɨ] and [o] were, in some subjects' opinion, closer to the phonetic prototypes of /i/ and /a/. If the responses of the two subjects favouring [i] over [ɨ] were disregarded, the number of misidentifications or [ɨ] as [i] would not exceed 2 for any of the stimuli.

Apart from the 145 cases in which [ɨ] was perceived as [i], there also occurred 38 responses classifying this vowel as [e]. This would contradict the hypothesis that the scaling factors for [ɨ] were too high. If that indeed had been the case, reactions of the [e] type should not have occurred, especially in view of the considerable perceptual distance between these two vowels (cf. e.g. [26]).

An analysis of misidentifications of [ɨ] as [e] indicates that such errors were particularly common if the fundamental frequencies of an utterance did not "fit" the formant frequencies or, strictly speaking, if it was too low in relation to them. Typical example of this kind are the stimuli in which female formants were combined with male FO or in which child's formants were combined with female FO.

A "misfit" of the same kind was also, at least partly, responsible for the considerable number of identification errors for [o]. Those stimuli in which the discrepancy between FO and formant frequencies was the largest were most often identified as [a]. For example, stimuli combining female formants and male FO as well as child's formants and female FO were classified as [a] by as many as 16 subjects.

Undoubtedly, both FO itself and the distance F1–FO play a role in the perception of vowel height (e.g. [3], [9], [27], [35], [36], [37]). TRAUNMÜLLER [36] has shown, for example, that the perceived vowel openness changes with FO, even if formant frequencies remain constant. CARLSON et al. [3] conclude that for a vowel to retain its phonetic identity an increase in FO must be accompanied by an appropriate increase in formant frequencies.

It has to be noted here that the misfit between FO and formant frequencies in some stimuli is a consequence of the very design of the present experiment. The objective was not to generate "prototypical" (i.e. easily identifiable) vowel tokens but stimuli in which various formant sets would be combined with various FO patterns.

Considering the "non-prototypical" and, perhaps, even unrealistic nature of some stimuli (e.g. female formants with male FO), identification **must** have provided difficulties, especially with non-quantal vowels and those perceptually similar to others (e.g. [ɨ] and [o]). The difficulty of the task was additionally increased by the variety of "voices" (no two stimuli in the whole set were identical) occurring in random order (cf. [21]). This manner of presentation forced the listeners to treat each stimulus separately, i.e. without referring it to the ones heard previously.

From the acoustic point of view, realisations of the same vowel by a male and child's voice are markedly different ([20]). In spite of that, listeners can, of course, identify such two utterances as representing the same phoneme ([38]). This perceptual equivalence is arrived at by way of voice normalisation, the mechanism of which is not yet fully understood. An immediate proof of the reality of this process in speech perception is a longer reaction time in the identification of vowels produced by a number of different speakers in comparison with a single speaker vowel identification task ([34]). This means that voice normalisation (so frequent in the present experiment) makes vowel identification more difficult and can decrease its effectiveness.

Probably due to the random order of stimulus presentation, no vowel contrast effect occurred in the responses (cf. e.g. [4], [30], [31])[5].

An interesting aspect of the problem of discrepancy between FO and formant frequencies in an utterance is highlighted by [8], [24] and [38]. It appears that a skilful insertion of one word spoken by a man imitating a high FO of a child's voice into a recording of a phrase spoken by a child results in a considerable decrease in the recognition score of the vowel(s) contained in this word. This primarily confirms the importance of phonetic context in speech sound perception: the preceding context provides a reference frame which, under natural communication conditions, facilitates (optimizes) identification of the speech sounds that follow ([38]). In the case of a "mystification" such as was used in the papers quoted, the lack of agreement between the acoustic structure of the male utterance and the perceptual reference frame was bound to cause a high identification error rate (ca. 54 per cent in [8]). It has to be noted that while the imitation of child's FO by the male speaker was quite successful, the difference in formant frequencies between realisations of the same vowel by the two speakers was considerable. In [38], for example, male [ɛ] was most similar to child's [œ] and male [œ] showed greatest similarity to child's [y]. As can be expected, even if stimuli of this kind were presented in isolation, the discrepancy between FO and formant frequencies would result in erroneous recognition of the (intended) vowels. Indeed, this was the case with some [o] stimuli in the present experiment. The reason why some female [o]s were recognized as /a/ was that their formant frequencies only slightly differed from the formant frequencies of male [a], which, in turn, were almost identical with those of child's [o] (cf. Tables

[5] With 3 sec ISI, this effect should, in principle, be negligible (cf. [7]).

4 and 7). In those cases where FO was too low, this inevitably led to a change of phonetic category.

The perceptual identity of vowels is commonly claimed to be determined by their formant frequencies, especially F1 and F2. It is also emphasized that, unlike consonant perception, the perception of vowels is non-categorical, which means that more vowels can be discriminated than identified ([6], [25], [30]). Modifying formant frequencies within cértain limits may, thus, lead to a change in perceived voice quality, but not necessarily to a change in phonetic category. Similar effects can be obtained by manipulating FO in an utterance with fixed formants ([36]). For a change in phonetic category to take place, the range of FO or formant shifts must exceed certain critical values (as can be expected, formant shifts play a more decisive role in this respect).

It follows from the above that vowel perception is determined not only by formant frequencies alone, but also by their relation to FO. This conclusion is contradicted by the findings of Summerfield and Haggard [34] who claim that while the significance of FO in perceptual identification of voices is beyond doubt, its role as a normalizing factor in vowel recognition is negligible. It has to be noted, however, that in the work quoted the difference in FO between voice I and the remaining ones amounted to just 20 Hz, which was probably too little for the normalizing "action" of this parameter to take place (all the four voices represented the same, i.e. male, voice category).

Van Bergem et al. [38] put forward a hypothesis according to which separate male, female and child's vowel templates exist in subjects' memory, formed on the basis of past language experience. Identification errors that occurred in their experiment are explained by the authors as resulting from the confusion of templates by the listeners.

No matter whether the above claim is valid or not, it seems certain that FO carries important information on voice type and, therefore, plays a crucial role in voice normalisation, thus making verbal communication more efficient. The results of perceptual voice categorisation, discussed below, support this supposition.

6.2. *Voice categorisation*

In order to ensure maximally objective conditions of the experiment and to avoid influencing the subjects' decision criteria in any way, no information was provided to the subjects as to the purpose of the experiment. In particular, they were not told that in the material to be presented they might come across high male voices or low female voices. Their decisions were thus fully independent in all cases.

Just as in vowel identification, response bias was observed in some listeners' answers, evidencing the relative nature of voice category "prototypes". Two subjects used the "woman" response as a reaction to stimuli judged by the majority to represent a male voice and one other consistently applied this label (i.e. "woman") in reference to stimuli predominantly classified as child's.

An analysis of the responses in Figures 1 . . . 6 indicates that the preceding context, which proved not to be significant in the vowel identification task, had some effect on voice categorisation results. As can be seen, some stimuli clearly differ from their immediate surrounding with respect to the type of voice category qualifiers ascribed to them. Both contrast and attraction effects were observed. If, for example, the difference between two successive stimuli was considerable and the former was judged as "definitely male", the subjects more uniformly classified the latter as representing a female voice (contrast effect). If, on the other hand, the preceding stimulus was considered male and the difference between this stimulus and the next was not great, the tendency prevailed for this following stimulus to be classified as male as well (attraction effect).

6.2.1. *Male voice and female voice.* Voice pitch was undoubtedly the principal factor determining the listeners' classification of utterances as representing male or female voices. This is evidenced by the fact that (1) high formant frequencies in themselves did not guarantee that the "woman" response would predominate and (2) even utterances with low (i.e. potentially male) formants but high FO were classified as female. In the data obtained, the "woman" responses begin to appear at mean FO value of 179 Hz, and prevail at mean FO equal to 192 Hz. It is in this frequency range that the (perceptual) boundary between the male and female pitch probably lies.

It has to be noted that even in the case of stimuli which, owing to a misfit between FO and formant frequencies, were misidentified by a number of subjects, the voice qualifier in those erroneous responses was predominantly the same as in the majority of the correct ones. This fact confirms both the dominant role of fundamental frequency in determining voice category and the effect of FO on vowel identification.

6.2.2. *Female voice and child's voice.* The perceptual boundary between the female and child's voices seems somewhat more fuzzy than this between female and male voices. This may have resulted from the indefiniteness of the child's voice category or the indefiniteness of the very notion of the "child". Whilst the voice of a 5-year-old child is relatively easy to identify, differences in the acoustic structure of utterances produced by a woman and a 13-year-old boy may not be great. Obviously, it is difficult to establish what definitions of the "child" were adopted by the individual listeners for the purposes of the present experiment.

The indefiniteness of the child's voice caused, among others, greater contrats and attraction effects than those noted for stimuli around the male-female boundary (cf. Figures 1. . .6). Also, the choice between the "woman" and "child" responses seemed to be affected to a greater extent by formant frequencies of the stimulus, although at high mean FO values the „child" response predominated irrespective of formant frequencies. For the combined data, the perceptual boundary between the female and the child's voice lies in the range of mean FO values between 230 Hz and 243 Hz.

Slight deviation from this pattern can be observed for [u], in which case the boundary is somewhat shifted towards higher frequencies.

7.Conclusions

Contrary to what is still sometimes assumed, the perceptual phonetic identity of a vowel is not determined solely by its formant frequencies: they can only be interpreted on the basis of the information supplied by fundamental frequency. For a vowel to be perceptually distinct (and easily identifiable), its formant frequencies must combine with *appropriate* FO. If FO is too low in relation to the formants, the perceived vowel becomes more open; if, on the other hand, FO is too high, the vowel is perceived as more close.

Establishing voice category seems an indispensable condition of correct vowel identification. In other words, in order to understand **what** has been said it is first necessary to know **which voice category** the speaker represents. As the results obtained show, the largely necessary factor determining the perceptual classification of a given voice as male, female or child's is fundamental frequency. This goes to prove that FO guides the process of vocal tract normalization and thus makes spoken communication more efficient.

References

[1] J. ALLEN, M. S. HUNNICUTT, D. H. KLATT, *From Text to Speech: The MITialk System*, Cambridge University Press, Cambridge 1987.

[2] *AUDLAB User's Guide*, Centre for Speech Technology Research, University of Edinburgh, Edinburgh 1987.

[3] R. CARLSON, G. FANT, B. GRANSTRÖM, *Two-formant models, pitch, and vowel perception*, in: G. Fant and M. A. A. Tatham (eds.): Auditory Analysis and Perception of Speech, Academic Press, London 1975, pp. 55–82.

[4] K. CENTMAYER, *Interrelations of vowel perception and linguistic context*, in: G. Fant and M.A.A. Tatham (eds.): Auditiory Analysis and Perception of Speech, Academic Press, London 1975, pp. 143–152.

[5] T. CHIBA, M. KAJIYAMA, *The Vowel – Its Nature and Structure*, Tokyo 1941.

[6] L. CHISTOVICH, *Central auditory processing of peripheral vowel spectra*, J. Acoust. Soc. Amer., 77, 789–805 (1985).

[7] R. G. CROWDER, *Decay of Auditory Memory in Vowel Discrimination*, J. Exptl. Psychol.: Learning, Memory and Cognition, 8, No 2, 153–162 (1982).

[8] D. DECHOVITZ, *Information conveyed by vowels: a confirmation*, Haskins Laboratories Status Report, SR – 51/52, 213–219 (1977).

[9] M.-G. DI BENEDETTO, *On vowel height: acoustic and perceptual representation by the fundamental and the first formant frequency*, Proceedings of the 11th International Congress of Phonetic Sciences in Tallin, vol. 5, 198–201 (1987).

[10] P. D. EIMAS, J. L. MILLER, P. W. JUSCZYK, *On infant speech perception and the acquisition of language*, in: S. Harnad (ed.): Categorical Perception: The Groundwork of Cognition, Academic Press, Cambridge 1987, pp. 161–195.

[11] J. L. ELMAN, J. L. MC CLELLAND, *Exploiting Lawful Variability in the Speech Wave*, in: J. S. Perkell,

D. H. Klatt (eds.): Invariance and Variability in Speech Processes, Lawrence Erlbaum Assoc., Hillsdale 1986, pp. 360–380.

[12] G. Fant, *On the predictability of formant levels and spectrum envelopes from formant frequencies*, in: M. Halle, H. G. Lunt, M. McLean (eds.): For Roman Jakobson, Mouton, The Hague 1956, pp. 109–120.

[13] G. Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague 1970 (2nd ed.).

[14] G. Fant, *A Note on Vocal Tract Size Factors and Nonuniform F-Pattern Scalings, in*: G. Fant: Speech Sounds and Features, The MIT Press, Cambridge, Massachusets, 1973, pp. 84–93.

[15] G. Fant, *Nonuniform Vowel Normalization*, Speech Transmission Laboratory QPSR, 2–3, 1–19 (1975).

[16] J. L. Flanagan, *Estimates of the Maximum Precision Necessary in Quantizing Certain "Dimensions" of Vowel Sounds*, J. Acoust. Soc. Amer., **29**, 533–534 (1957).

[17] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, Springer –Verlag, Berlin 1965.

[18] C. S. Hasek, S. Singh, *Acoustic attributes of preadolescent voices*, J. Acoust. Soc. Amer., **68**, 1262–1265 (1980).

[19] E. B. Holmberg, R. E. Hillman, J. S. Perkell, *Glottal airflow and pressure measurements for female and male speakers in soft, normal and loud voice*, paper presented at the 114th Meeting of Acoust. Soc. Amer., Florida 1987.

[20] J. N. Holmes, *Normalisation in Vowel Perception*, in: J. S. Perkell, D. H. Klatt (eds.): Invariance and Variability in Speech Processes, Lawrence Erlbaum Assoc., Hillsdale 1986, pp. 346–357.

[21] D. Kahn, *On the identifiability of isolated vowels*, UCLA Working Papers in Phonetics, **41**, 26–31 (1978).

[22] D. H. Klatt, *Software for a Cascade/Parallel Formant Synthesizer*, J. Acoust. Soc. Amer., **67**, 971–995 (1980).

[23] D. H. Klatt, L. C. Klatt, *Voice quality variations within and across female and male talkers: implications for speech analysis, synthesis and perception*, paper presented at the 114th Meeting of Acoust. Soc. Amer., Florida 1987.

[24] P. Ladefoged, P. Broadbent, *Information Conveyed by Vowels*, J. Acoust. Soc. Amer., **29**, 98–104 (1957).

[25] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, M. Studdert-Kennedy, *Perception of the Speech Code*, in: E. E. David, Jr., P. B. Denes (eds.): Human Communication: A Unified View, McGraw-Hill, New York 1972, pp. 13–50.

[26] P. Łobacz, G. Demenko, *Dependence of the preception of segmental features of the Polish vowels on the structure of the long-term lexical-phonetic memory* (in Polish), Reports of the Institute of Fundamental Technological Research 40, Warsaw 1983.

[27] W. Majewski, J. Zalewski, *The role of fundamental frequency in the process of perception of synthetic Polish speech signals* (in Polish), Reports of the Institute of Telecommunication and Acoustics, Technical University of Wrocław, **13**, 37–50, (1973).

[28] R. B. Monsen, A. M. Engebretson, *Study of variations in the male and female glottal wave*, J. Acoust. Soc. Amer., **62**, 981–993 (1977).

[29] J. S. Perkell, D. H. Klatt (eds.), *Invariance and Variability in Speech Processes*, Lawrence Erlbaum Assoc., Hillsdale 1986.

[30] B. H. Repp, A. F. Healy, R. G. Crowder, *Categories and Context in the Perception of Isolated Steady-State Vowels*, J. Exp. Psychol.: Human Perception and Performance, **5**, 1, 129–145 (1979).

[31] B. H. Repp, A. M. Liberman, *Phonetic category boundaries are flexible*, in: S. Harnad (ed.): Categorical Perception: The Groundwork of Cognition, Academic Press, Cambridge 1987, pp. 89–112.

[32] A. G. Samuel, *Phonetic Prototypes*, Perception and Psychophysics, **31**, 4, 307–314 (1982).

[33] K. N. Stevens, *The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data*, in: E. E. David, Jr., P. B. Denes (eds.): Human Communication: A Unified View, McGraw-Hill, New York 1972, pp. 51–66.

[34] A. Q. SUMMERFIELD, M. P. HAGGARD, *Vocal tract normalisation as demonstrated by reaction times*, in: G. Fant, M.A.A. Tatham (eds).: Auditory Analysis and Perception of Speech, Academic Press, London 1975, pp, 115–141.

[35] A. K. SYRDAL, *Aspects of a model of the auditory representation of American English vowels*, Speech Communication, **4**, 121–135 (1985).

[36] H. TRAUMÜLLER, *Perceptual dimension of openness in vowels*, J. Acoust. Soc. Amer., **69**, 1465–1475 (1981).

[37] H. TRAUNMÜLLER, F. LACERDA, *Perceptual relativity in identification of two-formant vowels*, Speech Communication, **6**, 143–157 (1987).

[38] D. R. VAN BERGEM, L. C. W. POLS, F. J. KOOPMANS-VAN-BEINUM, *Perceptual Normalization of the Vowels of a Man and a Child in Various Contexts*, Speech Communication, **7**, 1–20 (1988).

Received April 30, 1990