ARCHIVES OF ACOUSTICS
19, 2, 185–199(1994)

THE SYNTHESIS OF FEMALE VOICES USING A SOFTWARE SYNTHESIZER

M. OWSIANNY

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

The existing cascade-parallel software speech synthesizer modelled on the system of Dennis Klatt was further improved. A new, more universal periodic pulse source was added to it and the program environment of the main procedure of speech synthesis was made more useful. The usability of the improved synthesizer in the synthesis of female voices was tested. Natural vowels produced by 4 female and 2 male speakers were resynthesized and subjected to auditory evaluation by 10 listeners. The synthesizer was controlled using data derived from LPC analysis. The subjects' task was to identify the natural and synthetic vowels presented in random order, to indicate which of them sounded unnatural, to specify each speaker's gender and, in a subsequent test following familiarization with the voices, to give his or her name. The results obtained are indicative of high naturalness of the synthetic speech, reflecting personal voice features, and corroborate the usability of LPC in the extraction of parameters for the synthesis.

1. Introduction

Interspeaker differences are one of the main sources of variation the listener has to cope with in speech perception. The basis for eliminating this variation seems to be the ability to discriminate voice categories. Even though the productions of a given phone by a male, female and child's voice show considerable differences in the acoustic structure, listeners are capable of their recognition and classification. Explanation of the mechanism of perceptual normalization of the speaker's vocal tract, i.e. of eliminating personal voice characteristics, is of particular interest to scientists dealing with automatic speech recognition. It seems that speech synthesis can significantly contribute to the solution of this problem. Modelling of various aspects of interspeaker variation or even certain extralinguistic phenomena is becoming important in speech synthesis. It is also the way of improving the naturalness of the synthetic voice.

Not always has the synthesis of female voices been successful, mainly because of the difficulties in analyzing the signal. Formant frequencies are commonly maintained to be the indicator of phonetic differences. The high fundamental frequency of the female and child's voices makes it difficult to determine the location of formants in the

frequency scale. In many phones spoken by women, the first formant and the first harmonic are very close to each other. This makes it difficult or even impossible to accurately measure F1 and to separate in the spectrum the source-related effects from the transfer function effects. A more breathy quality of the female voice does not make analysis any easier. Also, the range of possible female voices is bordered by male voices at the one end and by children's voices at the other, which makes listeners' evaluation of "femaleness" of a voice more critical. The above difficulties were successfully surmounted by D. KLATT in 1980 [8]. Using his well known software cascade-parallel formant synthesizer, he produced natural sounding copies of a few vowels and of a short sentence produced by a female speaker. Other successful imitations of female voices were presented by FANT, GOBL, KARLSSON and LIN in 1987, HOLMES (1989) and by CARLSON, GRANSTROM and KARLSSON (1990) [6]. In 1990 D. KLATT and L. KLATT presented their new, improved synthesizer [9].

The aim of the present work was to improve an earlier version of the software cascade-parallel formant synthesizer modelled on the system of D. KLATT [8], so that it could be used to successfully synthesize female voices. The usability of the synthesizer for this purpose was examined in listening tests.

2. Description of the synthesizer

The software synthesizer SMOK was designed according to the standard set by the widely known system of Dennis KLATT [8]. Nevertheless, it is not a faithful copy of the

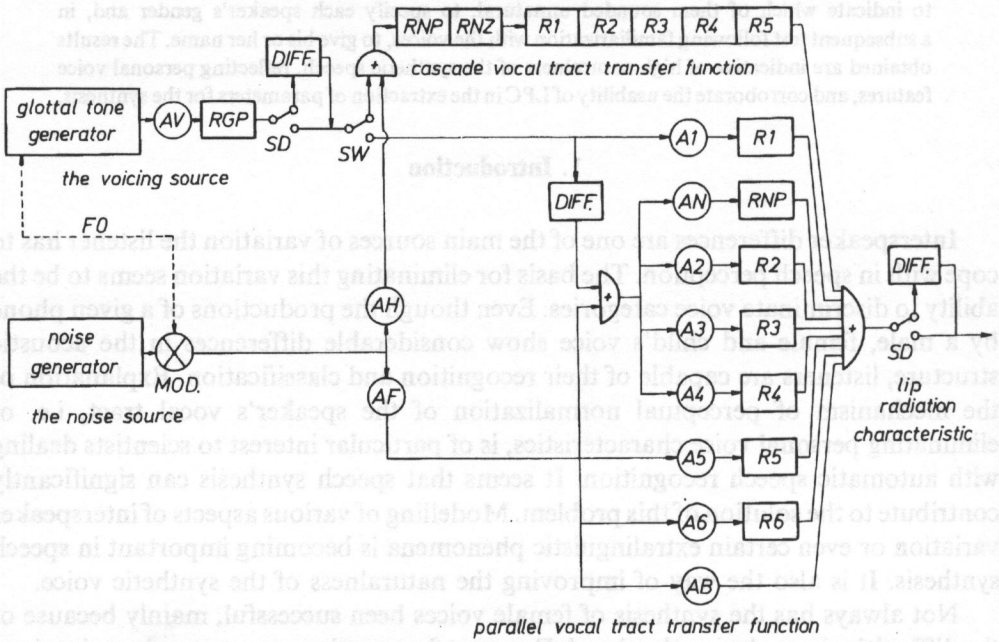

Fig. 1. Block diagram of the main procedure of the software cascade-parallel formant speech synthesizer ([8] with some modifications). R... — digital resonators, A... — amplitude control parameters, DIFF. — signal differentiation, MOD. — modulator, SD, SW — switches

synthesizer mentioned, even though a number of elements, and particularly the very idea of such a solution, are similar. The program was written in Pascal and implemented on an IBM PC equipped with a 12-bit D/A converter.

Figure 1 presents a block diagram of the SMOK synthesizer. It contains two excitation sources: a generator simulating the laryngeal voice source and a noise generator. Shaping the transfer function envelope of the signal is done by a set of digital filters connected in series or in parallel. The appropriate radiation characteristic is obtained by differentiating the signal output by the filters. The signal calculated in this way is then subjected to D/A conversion and amplified.

2.1. The excitation sources

The periodic pulse source has a decisive effect on the naturalness of the synthetic voice. In an earlier version of the synthesizer, the glottal source was simulated by an impulse train generator [8]. The impulses of the frequency of fundamental tone were filtered digitally, whereby an approximation to a glottal waveform was obtained. Each filter was defined by two parameters: the centre frequency and the bandwidth.

Even though this solution made it possible to synthesize voices of good quality, controlling the synthesizer proved troublesome. Filter parameters had little to do with features defining the source spectrum. Also, the naturalness of synthetic female voices required improvement. Therefore, following the example of D. KLATT and his daughter, who presented their new synthesizer in 1990 [9], this time not revealing details of its construction, the present author improved the voicing source model. Parameters controlling the new excitation tone also describe its spectrum, which is an important advantage. The waveform $U(t)$, corresponding to the open phase of the glottal excitation period or, more accurately, to the phase of abducting opening and adducting of the vocal folds, is calculated according to the formula suggested by ROSENBERG (1971) quoted in KLATT and KLATT [9]:

$$U(t) = at^2 - bt^3, \tag{1}$$

for a variable t defining the current time ($t \in \langle 0, a/b \rangle$), where a and b are constants whose values are defined by amplitude of voicing AV [dB] and by the open quotient (OQ [%]) of the glottal waveform, in per cent of a full period. These coefficients can easily be obtained by calculating the maximum of the function (1) and assuming that the value of the function at this point is equal to AV.

$$a = bOQ; \quad b = \frac{27\,AV}{(4\,OQ^3)}. \tag{2}$$

The closing phase of the glottal period is approximated by a straight line. The resulting waveform is passed through a low-pass digital filter (resonant frequency $F=0$) whose bandwidth BW [Hz] is a function of the parameter TL [dB] which defines the spectrum envelope fall-off above 3 kHz in relation to a typical spectrum envelope falling off at

-12 dB per octave and sampling frequency T. Using the formula for the sampled frequency response of the transfer function of a digital resonator (cf. [8], equation (3), p. 374), the following dependence was calculated:

$$BW = \frac{\ln(x+\sqrt{x^2-1})}{\pi T}; \quad x = \frac{1-TL\cos(2\pi 3000T)}{1-TL}. \tag{3}$$

The source parameters are related in such a way that the open quotient OQ affects the source spectrum in the low frequencies and has only a slight effect on spectrum amplitude in the higher frequencies. High frequency spectrum components are changed using the TL parameter. By selecting appropriate values of TL, OQ and AH (amplitude of aspiration/breathiness noise optionally superimposed on the glottal excitation signal, if considered desirable) it is possible to obtain good approximations to natural source spectra for various male and female speakers. A block diagram of the voicing source is presented in Fig. 2.

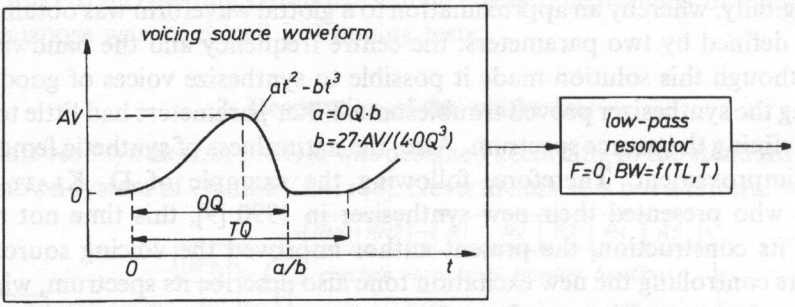

Fig. 2. Block diagram of source tone generation ([9] with some modifications).

In order to eliminate the monotonous voice quality resulting from the constant value of FO, a random period-to-period FO fluctuation was applied. A quasi-random number ΔFO, obtained by summing three slow-varying sinusoidal functions, was added to the nominal value of FO [9]

$$\Delta FO = \left(\frac{FL}{50}\right)\left(\frac{FO}{100}\right)[\sin(2\pi 12.7t)+\sin(2\pi 7.1t)+\sin(2\pi 4.7t)], \tag{4}$$

where t stands for the current time and FL is a parameter defining in per cent the level of the fundamental frequency modulation.

The glottal tone generator, controlled by means of the fundamental frequency FO, is the excitation source for voiced speech sounds, whereas the noise generator is used to obtain voiceless sounds or whispered speech. In the case of voiced fricative sounds, the noise is amplitude modulated with the frequency FO in the MOD modulator. The synthesizer can thus generate glottal tones of various shapes for voiced speech sounds, two kinds of aspiration noise: normal and amplitude modulated, and two similar kinds of affrication noise for affricate speech sounds.

2.2. Transfer function of the vocal tract

The transfer function of the vocal tract can be realized in two ways, i.e. by using cascade-parallel or just parallel configuration of the formant resonators. In the former, voiced and aspirated sounds are obtained in the cascade branch, whereas fricatives and affricates in the parallel one. If the operator wishes to independently control amplitudes of individual formants, he may use the parallel branch to generate all speech sounds. Since formants are the natural resonant frequencies of the vocal tract, and frequency locations are independent of source location, the formant frequencies of cascade and corresponding parallel resonators are identical.

The software cascade-parallel formant speech synthesizer described here is controlled by means of 38 parameters. 13 of them are constant throughout the utterance and are used to select the optimum configuration of the system. The remaining 25 parameters vary in time on a frame-to frame basis and define the shape of the glottal waveform as well as the vocal tract transfer function. Parameters are prepared in a specialized editor which makes it possible to input synthesis data in a textual or graphic form, display, modify, insert or clear them, as well as save or retrieve them. The effect of such operations can be seen on the monitor screen in the form of a glottal waveform, a vocal tract transfer function graph or the waveform of the synthetic output signal. Glottal tone and output signal waveform graphic presentations can be verified aurally. Detailed principles of controlling various parameters in the synthesis of English speech sounds can be found in [8, 9, 11, 12].

Starting the synthesis causes the spectral parameters referring to the consecutive frames of the signal to be transformed into a form used by the main procedure to control the digital filters. Amplitudes expressed in decibels are changed into a linear form, higher formant amplitudes are corrected, and digital resonator and antiresonator coefficients are calculated from formant frequencies F and bandwidths BW. These operations are carried out for each frame and the values obtained are sent to the main procedure, a schematic diagram of which is shown in Fig. 1.

2.3. Lip radiation characteristic

The acoustic pressure measured at the distance r from the lips is proportional to the temporal derivative of the lip-plus-nose volume velocity, and inversely proportional to r. The transformation is simulated in the synthesizer by differentiating the output signal.

2.4. Features of the program

The proper organization of the operating memory used by the program has a decisive effect on the maximum duration of the generated signal. Values of consecutive signal samples are stored on a heap. After starting the program on a typical IBM PC, the user has about 500 kB heap at his disposal. This means that a signal lasting 25 s can be handled at present, which is a sufficient duration considering the experimental-laboratory character of the synthesizer model presented.

The program is driven by an easy-to-use pull-down menu. It offers a number of useful features which make it functional and convenient to operate. They include, among others:

- the possibility of inputting the signal directly from a microphone connected via amplifier with an A/D converter, or from files stored in the external memory of the computer;
- concurrent editing of two signals (Fig. 3), e.g. a synthetic and a natural signal, combined with the possibility of producing acoustic output of the complete signals or their parts via the D/A converter and the amplifier;
- producing discrete spectra (DFT) at points indicated by the cursor, or contrastive spectra when two signals are being edited (see example in Fig. 4);
- the possibility of changing the sampling rate, combined with the automatic adjustment of the frequency of the anti-aliasing filter at the D/A converter output;

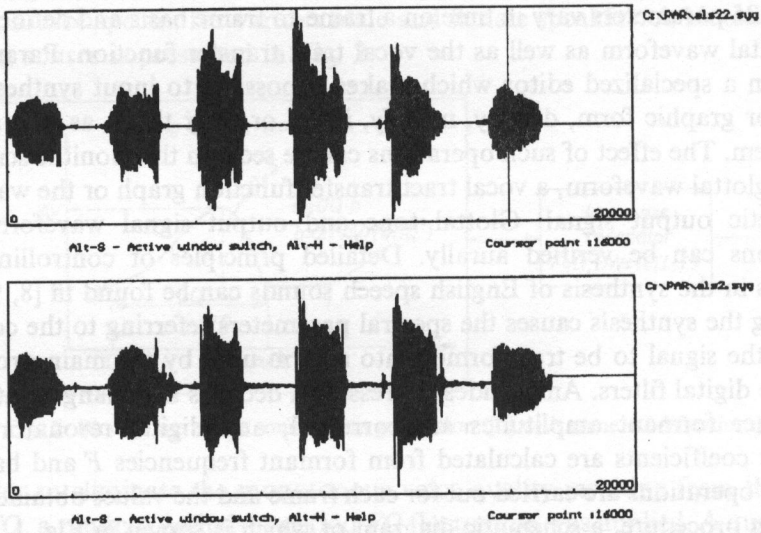

Fig. 3. Waveforms of two signals representing Polish vowels /i/, /i̇/, /e/, /a/, /o/, /u/ generated by the synthesizer (upper part) and their natural counterparts (lower part).

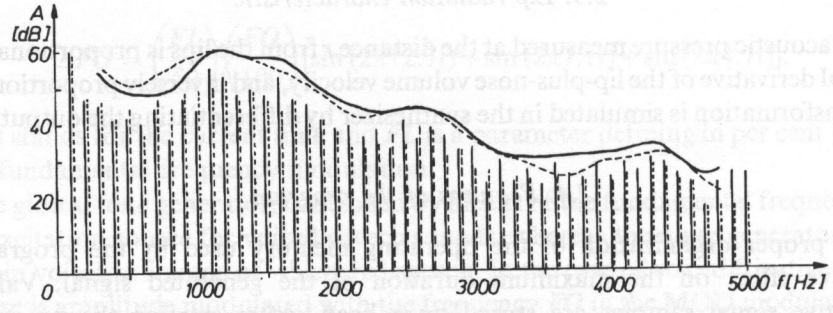

Fig. 4. Example of DFT spectra of a natural (dashed line) and synthetic (continuous line) vowel [a]. Horizontal lines express spectrum values at the indicated points on the frequency axis (every 100 Hz). The curves plotted above illustrate averaged versions of these spectra.
Natural [a] vowel produced by the female speaker AS.

• the possibility of a "joint" playback of a number of signals stored on disk (the total duration of the signal to be reproduced is limited by the size of the heap available).

3. Resynthesis of natural vowels

At the initial stage, recordings were made of six Polish vowels, viz. /i/, /ɨ/, /e/, /a/, /o/ and /u/, spoken in isolation one after another by 2 male and 4 female subjects. Average duration of each vowel was about 200 ms. Care was taken to ensure an optimum level of the recording and to prevent, as far as possible, intonation pattern variation. After low-pass filtering, sampling at 10 kHz and *D/A* conversion, the utterances were stored directly on the computer disk. Subsequently, suing an LPC-based software speech signal analyzer developed by P. DOMAGAŁA [3], parameters needed for the resynthesis of the vowels were extracted. An autocorrelation LPC method was applied in the analyzer, which realized the pole model of speech signal production and guaranteed stability of the prediction filter. For each frame, whose length was set to 128 samples, the program calculated the table of prediction filter coefficients, the table of reflection coefficients, the fundamental period measured in samples and the gain. Those parameters were used for the calculation of the poles of the transfer function by solving 10th order polynomial equations of real coefficients. Finding the poles made it possible to determine formant frequencies and bandwidths.

Parameters obtained in the way described, i.e. fundamental frequency *FO*, formant frequencies and bandwidths, gain, treated as the amplitude *AV* of the glottal excitation, as well as parameters constant for a given utterance, such as sampling rate, duration and the total number of frames were converted to a format accepted by the formant synthesizer SMOK and subsequently stored on the disk. Figure 5 demonstrates formant courses in six Polish vowels produced by a female speaker (AS).

Parameters extracted from the natural signal by means of LPC analysis could not be used directly in the synthesis. This referred particularly to formant bandwidths and, to a smaller degree, formant frequencies shown in Fig. 5. Analysis errors in some frames had to be corrected and spurious discontinuities smoothed. Fairly large

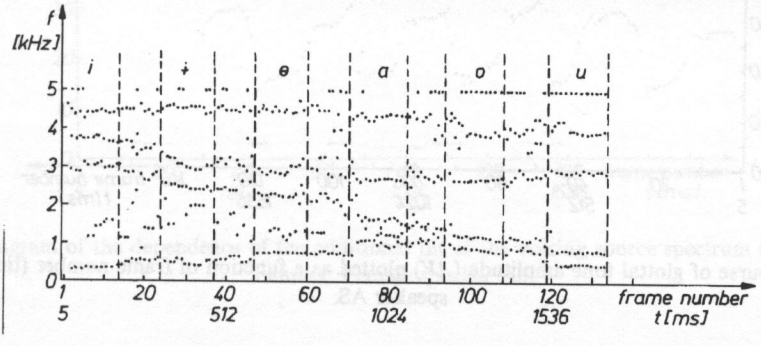

Fig. 5. Formant courses in six Polish vowels produced by a female speaker (AS). LPC analysis data.

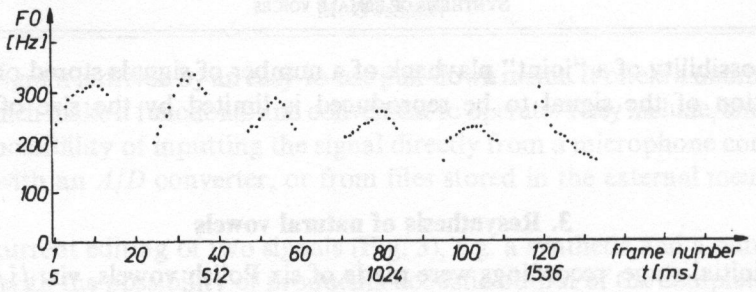

Fig. 6. The course of fundamental frequency (*FO*) plotted as a function of frame number (time). Female speaker AS.

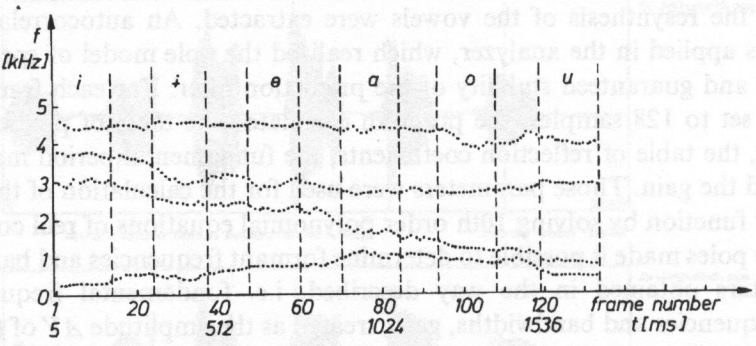

Fig. 7. Formant frequency courses in Polish vowels plotted as a function of frame number (time). Female speaker AS.

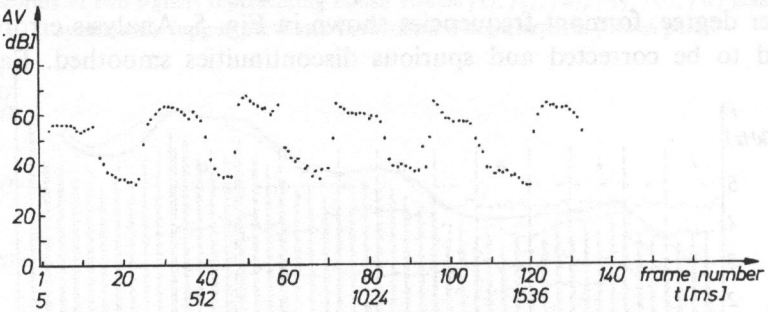

Fig. 8. The course of glottal tone amplitude (*AV*) plotted as a function of frame number (time). Female speaker AS.

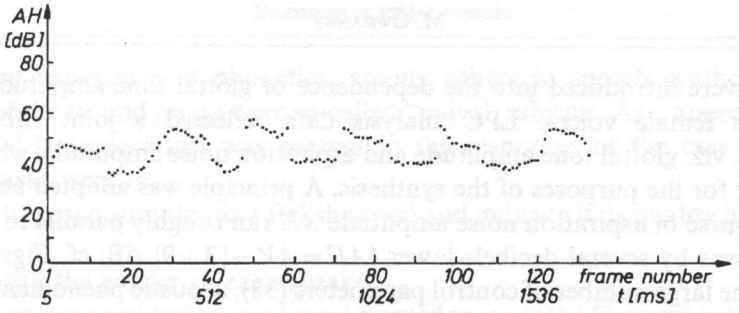

Fig. 9. The course of aspiration noise amplitude (*AH*) plotted as a function of frame number. Female speaker AS.

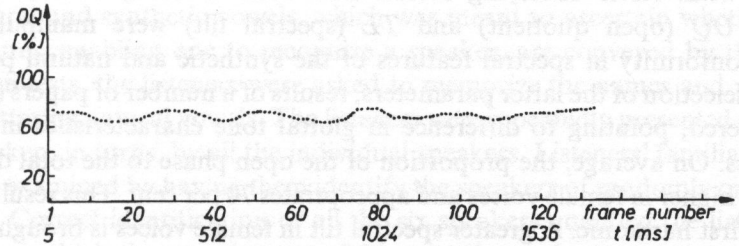

Fig. 10. The course of open quotient (*OQ*) plotted as a function of frame number. Female speaker AS.

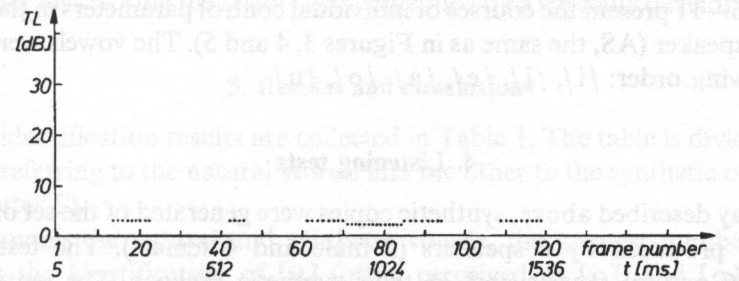

Fig. 11. Diagram of the dependence of the additional tilt of the voicing source spectrum (*TL*) on frame number. Female speaker AS.

corrections were introduced into the dependence of glottal tone amplitude on time, especially in female voices. LPC analysis data reflected a joint effect of two components, viz. glottal tone amplitude and aspiration noise amplitude, which had to be separated for the purposes of the synthesis. A principle was adopted according to which the course of aspiration noise amplitude *AH* ran roughly parallel to the course of *AV*, but was by several decibels lower ($AH = AV - \{3 \div 9\}$ dB; cf. Figs. 8 and 9).

Due to the large number of control parameters (38), acoustic phenomena observed in natural speech can be modelled with high accuracy. On the other hand, manipulating such a number of parameters requires an appropriate strategy. Therefore, in order to attain maximum naturalness of the vowels copied, certain principles of selection and optimization of parameters were adopted. Care was also taken to preserve, in the synthetic vowels, the personal voice features present in their natural counterparts. LPC data on the frequency (*FO*) and amplitude (*AV*) of the voicing source were adjusted first, whereupon formant frequencies and bandwidths were dealt with. After satisfying results had been attained, the voicing source parameters *OQ* (open quotient) and *TL* (spectral tilt) were manipulated until maximum conformity in spectral features of the synthetic and natural phones was reached. In selection of the latter parameters, results of a number of papers ([9, 10, 11]) were considered, pointing to difference in glottal tone characteristics in male and female voices. On average, the proportion of the open phase to the total duration of the period is higher in female voices and appropriates 70 per cent. This results in a high level of the first harmonic. A greater spectral tilt in female voices is brought about by the more symmetrical (sinusoidal) shape of the glottal pulse. In comparison with male voices, in which the spectral tilt approximates -12 dB per octave, spectrum fall-off in female voices is, on average, -15 to -18 dB/octave ($TL = 3..5$ dB). The first formant bandwidth in female voices is usually greater and comes near to 300 Hz. Also, female voices are typically more breathy, which was reflected by higher *AH* values (aspiration noise amplitude) applied (cf. [7, 9, 11]). With some voices, the overall effect was improved by introducing slight nasalization.

Figures 6–11 present the courses of individual control parameters in the utterance of a female speaker (AS, the same as in Figures 3, 4 and 5). The vowels were recorded in the following order: /i/, /ɨ/, /e/, /a/, /o/, /u/.

4. Listening tests

In the way described above, synthetic copies were generated of the set of six Polish oral vowels produced by 6 speakers (2 male and 4 female). The test materials contained 36 natural vowels and 36 their synthetic replicas. The materials were randomized, using a specially developed program, and recorded on a high quality Revox tape recorder. By repeating this procedure, three test sets were prepared, each containing 72 vowels, differing only in the order of the stimuli. The inter-stimulus interval was 4 s. Ten subjects (4 women and 6 men) divided into two equal groups participated in the listening tests. The first group was composed of persons with

professional experience in phonetics, among others in speech synthesis, and the members of the second group were so-called "naive" subjects. As it appeared later, no professional training effect was present in the responses of the two groups. The listeners' tasks were to:

1° identify each stimulus as a Polish vowel and indicate if its quality is synthetic or unnatural;

2° identify the gender of each speaker.

Listening test were conducted in a sound-treated room at the Department of Acoustics Phonetics, IFTR PAS, in Poznań. Each task was fulfilled separately. The listeners responded by filling in answer sheets. In three subsequent sessions, at least one day apart, the listeners carried out the first task. In each session, they were presented a test of a different (random) order of the stimuli. The second task was carried out once only, as practically no gender identification errors occurred in the first session.

Seven subjects (2 women and 5 men) took part in the second phase of the listening tests. Their task consisted in identifying each of the speakers' voices in a randomized set of natural and synthetic vowels, which was meant to ascertain whether personal voice features, enabling one to recognize a speaker, are conveyed by the synthesis. Prior to the tests, the listeners were asked to memorize the names and voices of the speakers uttering natural vowels. The listeners were repeatedly presented sequences of vowels spoken, in turns, by all the individual speakers. Listeners' familiarity with the voices was examined by having them identify the speakers of randomly ordered vowel sequences. Correct identification of all the six speakers qualified the listeners to the next stage at which the complete set of 36 natural vowels was presented to them in random order. Only three speaker identification errors were allowed. If the number of erroneous responses was greater than 3, the training phase was repeated. Three or more sessions were usually needed for a listener to reach a satisfying idenfitication score. Only then was it possible to administer the main test in which the speakers of all the 72 natural and synthetic vowels were to be recognized. Each listener participated in three test sessions. The whole process of familiarization with the voices, dialogue with the computer and storing the results was controlled by a specially developed program.

5. Results and conclusions

Vowel identification results are collected in Table 1. The table is divided into two parts, one referring to the natural vowels and the other to the synthetic ones (marked with the letter S).

Both among the natural and synthetic vowels, the greatest numbers of errors occured in the identification of [u] (often perceived as [o]) and [o] (frequently recognized as [a]). One of the reasons for that is the articulatory, acoustic and perceptual similarity of the vowels involved. An essential part is also played by context-related effects. The model natural vowels were produced, in a quasi-isolated way, in the order in which they appear in Table 1 (i.e. [i], [ɨ], [e], [a], [o], [u]). The pauses between them did not usually exceed 200 ms, which gave rise to co-articulation

Table 1. Vowel identification results. Synthetic vowels are marked with the letter S.

vowel presented	number of erroneous response	error rate
i	2 (y)	1.1
ɨ	7 (e), 1 (i)	5
e	–	0
a	3 (o)	1.9
o	17 (a)	10.6
u	22 (o)	13.8
iS	4 (y)	2.5
ɨS	4 (e)	2.5
eS	–	0
aS	6 (o)	3.7
oS	28 (a)	17.5
uS	37 (o), 2 (x)	24.4

effects. This can be observed in Figs. 5 and 7 which show formant courses in the six vowels. Needless to say, co-articulation made vowel identification more difficult. Also, the variety of the (male and female) voices occurring in random order did not make the listeners' task any easier: each stimulus had to be dealt with separately, i.e. without being referred to the ones heard previously. Analysis of responses of the individual listeners points to a certain responses bias. Three of the ten subjects made, on average, 26 identification errors each in three sessions, the average for the remaining seven subjects being 7 errors, i.e. 2–3 errors per session. For the synthetic vowels, the error rate is about twice as high. This is no surprise considering that the synthetic vowels were generated as close copies of their natural originals, even if the latter had not been spoken very carefully or had other deficiencies.

Table 2 presents evaluation to vowels with unnatural quality, obtained in three listening sessions from the group of experienced phoneticians (Table 2.1) and the group of naive listeners (Table 2.2).

Four complementary response categories were distinguished:

Table 2. Results of vowel quality evaluation by phoneticians (Table 2.1) and naive listeners (Table 2.2)
Table 2.1 description in the text.

	PD			HK			IN		
	1	2	3	1	2	3	1	2	3
S→S	16	14	15	11	14	9	18	16	22
S→N	20	22	21	25	22	27	18	20	14
N→S	7	6	5	5	2	3	16	14	17
N→N	29	30	31	31	34	33	20	22	19
	LR			BS					
	1	2	3	1	2	3			
S→S	11	14	16	19	22	26			
S→N	25	22	20	17	14	10			
N→S	2	12	7	10	11	15			
N→N	34	24	29	26	25	21			

Table 2.2.

	AG			TH			MK		
	1	2	3	1	2	3	1	2	3
S→S	17	15	17	22	18	13	15	10	16
S→N	19	21	19	14	18	23	21	26	20
N→S	13	14	8	10	13	10	10	10	17
N→N	23	22	28	26	23	26	26	26	19
	KO			IP					
	1	2	3	1	2	3			
S→S	25	23	21	13	18	20			
S→N	11	13	15	23	18	16			
N→S	9	4	7	6	6	6			
N→N	27	32	29	30	30	3			

- synthetic vowel judged to be a synthetic vowel (S→S);
- synthetic vowel judged to be a natural vowel (S→N);
- natural vowel judged to be a synthetic vowel (N→S);
- *natural vowel judged to be a natural vowel (N→N).*

In order to ascertain whether the differences in responses were significant statistically, a chi-square test was applied:

$$\chi^2 = \Sigma (f_i - f_t)^2, \tag{5}$$

where f_i denotes experimental sample size and f_t stands for expected (theoretical) size. The probability of a single event, i.e. of the identification of a synthetic vowel as a synthetic or a natural vowel is equal to 0.5, which, for the test containing 36 synthetic vowels and the same number of natural ones, gives the f_t value equal to 18. According to the null hypothesis adopted, the distribution of the listener's responses and the distribution expected are equivalent, any differences being due to chance. The χ^2 values were calculated separately for each series and for each listener. Only the first two categories (S→S and S→N) were considered, as the main problem investigated was the identifiability of stimuli in the set presented. It was found that at the confidence level $\alpha = 0.05$ and the number of degrees of freedom $df = 1$, the null hypothesis was true if the S→S category size was smaller than or equal to 23. This meant that at the confidence level adopted the listener did not identify synthetic vowels. At the same time, if the N→S category size was greater than 23, the result of the test would have to be discarded because of the listener's bias favouring the "synthetic quality" response. It can easily be seen that, of the ten listeners examined three times each, only in two cases were the synthetic vowels identified correctly. This bears evidence of the high quality of the vowels synthesized.

The other task of the listeners' consisted in identifying speakers' gender. Tests used earlier, each comprising 72 randomly ordered natural and synthetic vowels represen-

ting male and female voices, were applied for that purpose. Ten listeners participated in the tests. Seven of them did not make any errors. The few errors found among the remaining three subjects' responses were probably due to chance. As it became apparent that speaker gender identification did not pose any problems both in the case of the natural and synthetic vowels, the test was administered once only.

The third test was the most difficult one both for the listeners, from whom its complicated scenario required considerable concentration, and for the synthesizer, which was not only to produce natural sounding vowels but also to preserve in them the personal voice characteristics of the original speakers. The synthesizer seems to have done quite well. Appropriate results are shown in Table 3.

The mean number of errors made by the listeners in the identification of speakers of the natural v. synthetic vowels is 3.5 to 9.0. The number of speaker identification errors obtained for the natural vowels is somewhat higher than the threshold value adopted in the preliminary tests (3 errors). This is a natural consequence of including synthetic vowels in the stimulus set. Speaker identification rate obtained for the synthetic vowels is over twice as low as that for the natural ones. This is not a negative

Table 3. Mean numbers of errors committed by the individual listeners in the identification of speakers of the model natural vowels and their synthetic copies. Pooled results of three listening sessions are shown.

	BS	PD	JI	IN	MO	GD	LR
Nat.	2.7	4.3	3	2	2.7	4.7	4.7
Synt.	9	8.7	9.7	8	7	9.3	11.3

result considering that if the listeners had not identified the speakers and given chance responses, the number of errors would have approximated the mean, i.e. 30 (the probability of an error was equal to 5/6).

In the identification of the speakers, the listeners used various, often subjective criteria which they applied to the natural and synthetic utterances in the same degree. Similarities and differences occurring between the individual voices were reflected in their responses. It was considered interesting to examine relations among the natural voices and the corresponding relations among the synthetic ones. Classification of the voices was carried out using cluster analysis. For that purpose, two distance (dissimilarity) matrices corresponding to the natural and synthetic voices were created, the size of each of which was equal to the number of voices analyzed, i.e. 6. The lines corresponded to the voices presented and the columns to the voices received. The elements of the matrix were the numbers of the listeners' responses, normalized to the value of 1, defining the rate of dissimilarity between the voice presented and received. The value of "1" occurring in a matrix cell denotes that the two voices are dissimilar, i.e. that no listener had mistaken one for the other. The values along the main diagonal were of course zero.

Figures 12a and 12b present cluster diagrams, in the form of a hierarchical tree, obtained from distance matrix analysis. The StatSoft CSS: Statistica software packet was used to this end. A standard method of defining the distance between two clusters,

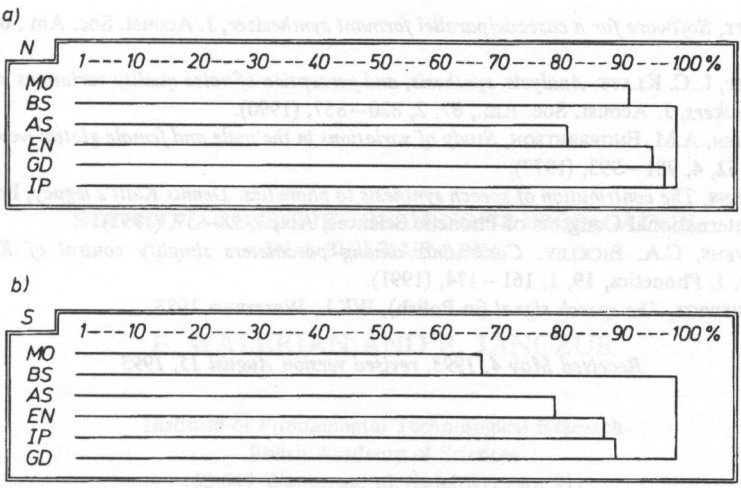

Fig. 12. Hierarchical trees defining dependences between individual objects for the natural (Fig. 10a) and synthetic voices (Fig. 10b). The voices are denoted by the speaker's initials: MO, BS (male), AS, EN, GD, IP (female).

the so-called "nearest neighbour" method, was applied. The diagrams begin at the left side where each object makes its own cluster. Further to the right, the voices group. The horizontal axis is scaled in per cent and defines the distance at which given clusters join. Thus, it provides a dissimilarity measure.

For the natural and synthetic voices, dependences among objects are similar. Female voices group in one cluster and are separated from the male ones. The male voices in their natural version are more dissimilar than their synthetic counterparts. Synthetic copies of the two most similar natural female voices (AS and EN) retain the high similarity of their originals.

The results obtained indicate that the new improved version of the software cascade-parallel speech synthesizer described can be used to generate high quality speech, and especially to imitate natural female voices.

References

[1] Cz. BASZTURA, *Sources, signals and acoustic patterns* (in Polish), WKŁ, Warszawa 1988.

[2] R. CARLSON, *Synthesis: modelling variability and constraints*, J. Phonetics, **19**, 1, 1–9, (1991).

[3] P. DOMAGAŁA, *Multidimensional statistical analysis of LPC parameters of phonetic segments in typical combinations* (in Polish), IFTR Reports, 21, (1988).

[4] J. IMIOŁCZYK, *Determination of perceptual boundaries between the male female child's voices in isolated synthetic polish vowels*, Archives of Acoustics, **16**, 2, 305–323, (1991).

[5] W. JASSEM, *The bases of acoustics phonetics* (in Polish), PWN, Warszawa 1973.

[6] I. KARLSSON, *Female voices in speech synthesis*, J. Phonetics, **19**, 1, 111–120, (1991).

[7] I. KARLSSON, *Evaluations of acoustic differences between male and female voices: a pilot study*, Speech Transmission Laboratory, Quarterly Progress and Status Report, **1**, (1992).

[8] D.H. KLATT, *Software for a cascade/parallel formant synthesizer*, J. Acoust. Soc. Am., **67**, 971–995, (1980).

[9] D.H. KLATT, L.C. KLATT, *Analysis, synthesis, and perception of voice quality variations among female and male talkers*, J. Acoust. Soc. Am., **87**, 2, 820–857, (1990).

[10] R.B. MONSEN, A.M. ENGEBRETSON, *Study of variations in the male and female glottal wave*, J. Acoust. Soc. Am., **62**, 4, 981–993, (1977).

[11] K.N. STEVENS, *The contribution of speech synthesis to phonetics: Dennis Klatt's legacy*, Proceedings of the 12th International Congress of Phonetic Sciences, Aix, **1**, 28–37, (1991).

[12] K.N. STEVENS, C.A. BICKLEY, *Constraints among parameters simplify control of Klatt formant synthesizer*, J. Phonetics, **19**, 1, 161–174, (1991).

[13] R. TADEUSIEWICZ, *The speech signal* (in Polish), WKŁ, Warszawa 1988.

Received May 4, 1993, revised version August 15, 1993