ARCHIVES OF ACOUSTICS
19, 2, 161–172(1994)

HIGH INTELLIGIBILITY TEXT-TO-SPEECH SYNTHESIS FOR POLISH[1]

J. IMIOŁCZYK, I. NOWAK AND G. DEMENKO

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

The paper presents the system of automatic synthesis of the Polish speech signal from text, developed over the last three years at the Department of Acoustic Phonetics in Poznań. The element generating the acoustic signal is a special-purpose IC controlled from a PC AT by means of original software comprizing the modules for text editing, phonemic transcription and synthesis of digital parameters. The speech signal, produced in real time, is highly intelligible, which opens up before the system a prospect of concrete applications in man–man and machine–man communication.

1. Introduction

In the period of rapid development of speech science which has taken place in the last three decades, speech synthesis has attracted attention of numerous research centres in many countries, notably United States, Japan, Sweden, France, and Great Britain. One of the most ambitious goals has been to create text-to-speech (TTS) systems which would automatically convert orthographic text (e.g. input from the computer keyboard) into the corresponding acoustic signal. Of such systems already in existence, the most successful one is probably MITalk [1]. Its operation is based on a great number of complex rules for text analysis and speech synthesis, the preparation of which took over 15 years.

In Poland, the first attempts to synthesize speech from text were made in the late 1970's [6]. Research continued throughout the 1980's (see e.g. [8]).

The principal objective of the present work, started in the late 80's, was to develop a TTS system for Polish which would be capable of producing **highly intelligible** speech. Since the system was to work in real time, no text analysis modules were included into it, which would be indispensable (although probably insufficient) for attaining truly natural sounding speech.

[1] This research was carried out within the IFTR project No 412

In the system presented, the so-called formant method was used, based on the "source-filters" model. The speech signal is generated by an IC from frequency components forming energy maxima (formants). With respect to the minimum speech element adopted, the synthesis can be referred to as allophone synthesis: each Polish phoneme is represented by a number of its contextual synthetic allophones which are combined in accordance with a set of concatenation rules to form longer utterances.

The operation of the system, programmed in Turbo Pascal for an IBM PC AT, embraces the following stages:

1. Input of an orthographic text using an inbuilt editor (keyboard, floppy disk).
2. Phonemic transcription.
3. Processing of allophones on the basis of segmental (modelling transients) and suprasegmental (duration, intensity and fundamental frequency) rules, using the information on the position occupied by the phones corresponding to those allophones within a word, clause, and sentence.
4. Superimposition of intonation contours on parameter value sequences.
5. Synthesis of the acoustic signal.

Transcription is executed for the **whole** text, whereas actions described in points 3..5 proceed in a sentence cycle until the generation of the last sentence of the text.

2. Basic data on the synthesizer

The element generating the acoustic speech signal is PCF 8200, an IC acting as a bank of five cascaded filters with programmable formant frequencies and bandwidths. The filters can be excited by a periodic pulse source or, alternatively, by a noise generator. Before being sent to the filters, the excitation signal is appropriately amplified. The filtered signal is subjected to D/A conversion using an 11-bit converter, and low-pass filtered below 5 kHz.

Parameter update interval (frame duration) of the synthesizer can be varied. It is defined as the product of the so-called **standard frame** duration (supportable values are 8.8, 10.4, 12.8 and 17.6 ms) and its multiplication factor (1, 2, 3 or 5). The ultimate frame duration (ranging from 8.8 * 1 up to 17.6 * 5) determines the rate of stepwise interpolation between the consecutive declared values of amplitude as well as formant frequencies and bandwidths. Fundamental frequency values are updated every 1/8 of the **standard** frame.

3. Phonemic transcription

Phonemic transcription is an indispensable element of any TTS system. Its function is to convert an orthographic text into a corresponding phonemic text — a sequence of conventional signs (symbols of phonemes) in which every consecutive sign corresponds to one speech sound (phone) or appropriate pause.

The algorithm of phonemic transcription for Polish used in the system is based on [16]. Its detailed description can be found in [10].

4. Segmental rules

As already mentioned, each Polish phoneme is represented in the system by a number of its contextual synthetic allophones, sufficient to ensure good quality of output speech. Depending on the phoneme, this number varies from 1 (e.g. for vowels) to 8 (for /p/ and /r/). The complete set consists of 82 allophones. In a vast majority of cases in which a phoneme was represented by more than just one allophone, the sole factor necessitating this distinction was the effect of the following context. Twelve types of it have been distinguished:

1. front vowel or [j]
2. back vowel or [w]
3. [i] or [j]
4. [i̯] 5. [e] 6. [a] 7. [o]
8. [u] or [w]
9. a fricative or an affricate
10. a voiceless consonant
11. a pause
12. all remaining contexts (except those specifically stated).

Only for a few phonemes was the necessity to distinguish allophones brought about by the effect of the preceding context. The following types of this context have been considered:

1. a pause
2. a voiceless consonant: utterance-medially, not before a voiceless consonant
3. a voiceless consonant: utterance-medially before a voiceless consonant
4. any non-voiceless phone in an utterance-medial position
5. a back vowel or [w] before a pause
6. a front vowel or [j] before a pause
7. a voiceless consonant before a pause
8. any other phone (except those specifically stated) before a pause.

Distinguishing 82 contextual allophones of phonemes only partly solved the problem of appropriate transient modelling. In order to remove all undesirable parameter discontinuities at phone boundaries, it was necessary to work out an extensive set of detailed rules which would determine, for any given phone combination, the duration of the transient and the values of control parameters within it. Amplitude, *F*1, *F*2, *F*3, *B*1 and temporal parameters have been subjected to modelling by means of rules. Typical examples are:

1. In the first 30-ms segment of a vowel following an [n] B1 is to be widened to 300 Hz.
2. The amplitude in the last frame of an [l] followed by an [u] is to be decreased by 2 scale units.
3. F3 in the first frame of an [i] following an [m] is to be lowered by 200 Hz etc.

As stated above, frame duration determines interpolation rate between the consecutive parameter values declared. By manipulating frame duration it is thus

possible to make a transition abrupt or, conversely, gradual and smooth, depending on concrete requirements. Whilst, no doubt, duration is principally a suprasegmental feature, it also has some bearing on phone distinctness, especially as far as duration of transitions is concerned. The system uses frames ranging from 8.8 ms (e.g. in the transition from [v] to [j]) to 64.0 ms (e.g. in the transition from [ɲ] to [a])

5. Suprasegmental rules

5.1. Duration rules

As a departure point, existing data were used on Polish vowel and consonant duration in one- and two-syllable nonsense words ([2] and [12], respectively) as well as on the rhythmical structure of Polish utterances ([13], [14]).

Each of the 82 synthetic allophones was assigned a standard duration, appropriate for as many contexts and positions in the utterances as possible[2]. Sufficiently detailed rules were to ensure proper duration of each phone in all other cases.

In order to supplement the available data, a DSP Sonograph 5500 was used, which made it possible to analyse the speech signal in real time and to make accurate measurements of phone durations. It was those measurements that served as the basis for generalization concerning temporal phenomena, subsequently formulated as rules.

The following factors affecting the temporal structure of Polish utterances have been considered in the rules:

I. *Phone type*

1. vowel 2. consonant.

Standard and minimum durations of phones varied. Among the oral vowels, [a] was the longest and [i̯] the shortest; among consonants, voiceless fricatives were several times as long as [r].

II. *Type of adjoining phones*

1.1. within a word / 1.2. across word boundary

a) vowel + a different vowel

b) vowel + identical vowel

c) [j], [w] or soft consonants ([ɕ], [ʑ], [tɕ], [dʑ], [c], [ɟ], [ɲ]) + vowel

d) vowel + [w], [j], [ɲ] or [ŋ]

e) voiced stop consonants or voiceless fricative consonants (except soft ones) as well as voiceless affricates (except soft ones) + vowel

f) consonant + [j]

g) consonant + [w].

[2] Absolute durations are obviously dependent on speech rate. For the purpose of synthesis, a moderate tempo was adopted.

The type of adjoining phones affects the duration of transition between their steady states. The initial portion of transition can be ascribed to the preceding phone, whilst the final portion to the following one. Thus, the overall duration of both phones is equal to the sum of their steady states and the transient. The longest transients occur in combinations of [j] and [ɲ] with vowels, especially back vowels [a], [o], [u], whereas the shortest — in combinations of bilabials with [j].

The effect of the type of phone combination on duration is to some degree dependent on whether the phones co-occur in a word (e.g. "g**EO**grafia" [geo'grafja] — "geography") or are separated by a word boundary (e.g. "ż**E O**grabia" [ʐe o'grabja] — "that (he) robs").

III. *Type of consonantal context following the vowel*

1. presence of voice / absence of voice
2. manner of articulation
3. single consonant / consonant cluster

As is generally known, vowels tend to be longer before voiced than before voiceless consonants. They are also somewhat shorter when followed by a consonantal cluster. The effect of the manner of articulation of a consonant on the duration of the preceding vowel is more complex: on average, vowels are shortest before stops and longest before [r].

IV. *Type of consonantal cluster*

1. type of consonants occurring in the cluster
2. length (number of elements) of the cluster

The degree of reduction (shortening) of consonants in a cluster is positively correlated with their intrinsic duration as well as with the number of elements of the cluster. Voiceless fricatives are, therefore, reduced most and [r] is not shortened at all (cf. I.2 above). [s] in "**wstr**ętny" (['fstrentny] — "abominable") is in turn shorter than in "**st**ąpać" ([stompać] — "to pace") because of the greater length of the first cluster.

Reduction of consonant duration in a cluster can also be brought about by the affinity of adjoining phones with respect to the place of articulation and voicing. For example, the reduction of closure after [m] is greater with [b] (both phones are bilabial and voiced, e.g. [uzembjeɲe] — "dentition") than with [p] (both phones are bilabial but differ with respect to voicing, e.g. [potempjeɲe] — "condemnation").

V. *Position/status of the syllable in the utterance*

1. utterance initial position
2. anacrusis (unaccented syllables utterance initially)
3. syllable with word stress
4. syllable with nuclear accent
5. utterance final position with a consonant as the last element

6. utterance final position with a vowel as the last element
7. unaccented syllable — apart from 2, 5 and 6 above.

Position and/or status of the syllable mainly affect the duration of vowels which, depending on factors involved, can vary by up to 100 ms. Vowels are shortest in anacrusis and in unaccented syllables and longest in utterance final position if not checked by a consonant.

VI. *Number of syllables in a foot*

Phone duration is negatively correlated with the number of syllables co-occurring in a rhythmical foot.

VII. *Type of pause*

The type of pause depends on the kind of punctuation mark generating it, e.g. a pause before a comma is shorter than before a full stop.

Distinguishing the above factors and determining their effect on phone length led to the formulation of duration rules. In a majority of cases, phone duration is co-determined by a number of factors. For example, the length of [e] in the utterance "on tu jest" ([on tu jest] — "he is here") is a result of joint effect of 8 factors: I.1, II.1.1.c, III.1, III.2, III.3, V.4, V.5 and VI. The set of duration rules in its ultimate form includes 300 single phone lengthening and shortening instructions. In the case of the vowel [e] in the example cited above ("on tu jest"), the instructions bring about the following effects:

1. standard duration of [e]: 70 ms (factor I.1)
2. long transient after [j]: +27 ms (factor II.1.1.c)
3. lengthening in a sentence-accent position: +97 ms (factor V.4)
4. shortening before a voiceless consonant: −18 ms (factor III.1)
5. shortening before a fricative: −18 ms (factor III.2)
6. shortening before a consonant cluster: −9 ms (factor III.3)
7. shortening in utterance-final position, not immediately before a pause: −25 ms (factor V.5)
8. one-syllable foot: no change (factor VI).

5.2. Rules for shaping intonation contours

5.2.1. *Modelling intonation in speech synthesis.* In want of data on Polish intonation, rules for controlling fundamental frequency were formulated on the basis of results obtained for other languages ([4], [5], [7], [9], [11]) and on current experimenting. The model put forward by FUJISAKI [3] has been adopted. In this model, accent components, calculated for individual accent groups, are superimposed on the so-called phrase component, determining baseline or declination of the utterance. The accent group is defined as a prosodic entity composed of an accented syllable and the following unaccented syllables [17]. Declination is defined by the function:

$$Gp(t) = Ap\alpha t \exp(-\alpha t) \tag{1}$$

where Ap denotes amplification coefficient, α is an attenuation coefficient and t stands for time. The accent component function Ga is expressed by the dependence:

$$Ga(t) = Aa\,(1-(1+\beta\,t)\,\exp(-\beta\,t) \qquad (2)$$

where Aa denotes accent amplification coefficient, β is an attenuation coefficient and t is time.

In order to adapt the model for Polish, an analysis was carried out of fundamental frequency contours in a newspaper text read three times by six speakers. On this basis, preliminary value ranges of control parameters Ap, α, Aa and β were determined. After their verification in perceptual tests, sets of tabulated phrase and accent component function values have been worked out for them.

5.2.2. Information necessary for FO control. It was considered necessary that the program generating intonation contours should make use of the following language data:

1) Sentence level data: number of clauses, sentence type (declaration, specific question etc.).
2) Data on current clause:
   a) position within the sentence.

The first and the last clauses are of particular importance, as they define sentence type and determine the dynamics of FO changes.

   b) type of ending: punctuation marks such as ( ), . ? etc.
   c) approximate duration in [ms].
   d) number of accent groups.

The number of accent groups affects the manner of FO contour modelling in the Fmax$-F$min variation range adopted.

   e) position of accent groups.

Modelling of FO contour in the first and the last accent group is particularly important, as it considerably affects the perception of intonation of the whole clause.

   f) structure of accent groups.

The selection of shape and dynamics of the intonation movement is to some extent dependent on the number and position of unaccented syllables.

3) Syllable level data.

Distinction of accented and unaccented syllables is essential for the determination of distances between consecutive FO maxima.

5.2.3. Rules for controlling phrase and accent components.

I. Phrase component control.

For the amplification and attenuation coefficients, the following value ranges were adopted:

$$Ap\text{: } 0.018-0.633 \quad \text{and} \quad \alpha\text{: } 1.14-8.00.$$

On this basis, 21 tabulated phrase function value sets have been worked out (7 clause duration classes combined with three FO levels — high, mid and low), with

starting values ranging from 100 to 124 Hz. For example, the value of *Ap* is maximum if a given clause is the first clause in a sentence, has the duration of more than 6.5 ms and begins with an accented syllable.

II. Accent component control.

On the basis of the adopted ranges of amplification and attenuation coefficients *Aa* and *β*, a set of 42 tabulated accent movements has been worked out (14 amplification values, covering the range from 6 to 84 Hz, combined with 3 attenuation values, corresponding to the **slow**, **fast** and **very fast** rate of FO change).

Three positions of accent groups within a clause, viz. **initial**, **medial** and **final**, have been distinguished. Also, three positions within a stressed vowel (i.e. timing) of intonation peaks corresponding to syllable accents have been adopted. The three timing-related types of peaks are:

- **early** peaks, occurring at the beginning of the stressed vowel
- **late** peaks, located around the middle of the stressed vowel
- **very late** peaks, occurring towards the end of the stressed vowel or even at the beginning of the following phone.

For the **initial** accent group, the following rules have been implemented:

1) Peak position in the stressed vowel: **very late**.
2) The rate of accent movements: **fast**.
3) The starting value of FO and its course conditioned by the presence/absence of anacrusis and voiceless consonant(s) at the beginning of the utterance.

FO trajectory in **medial** accent groups is characterized by:

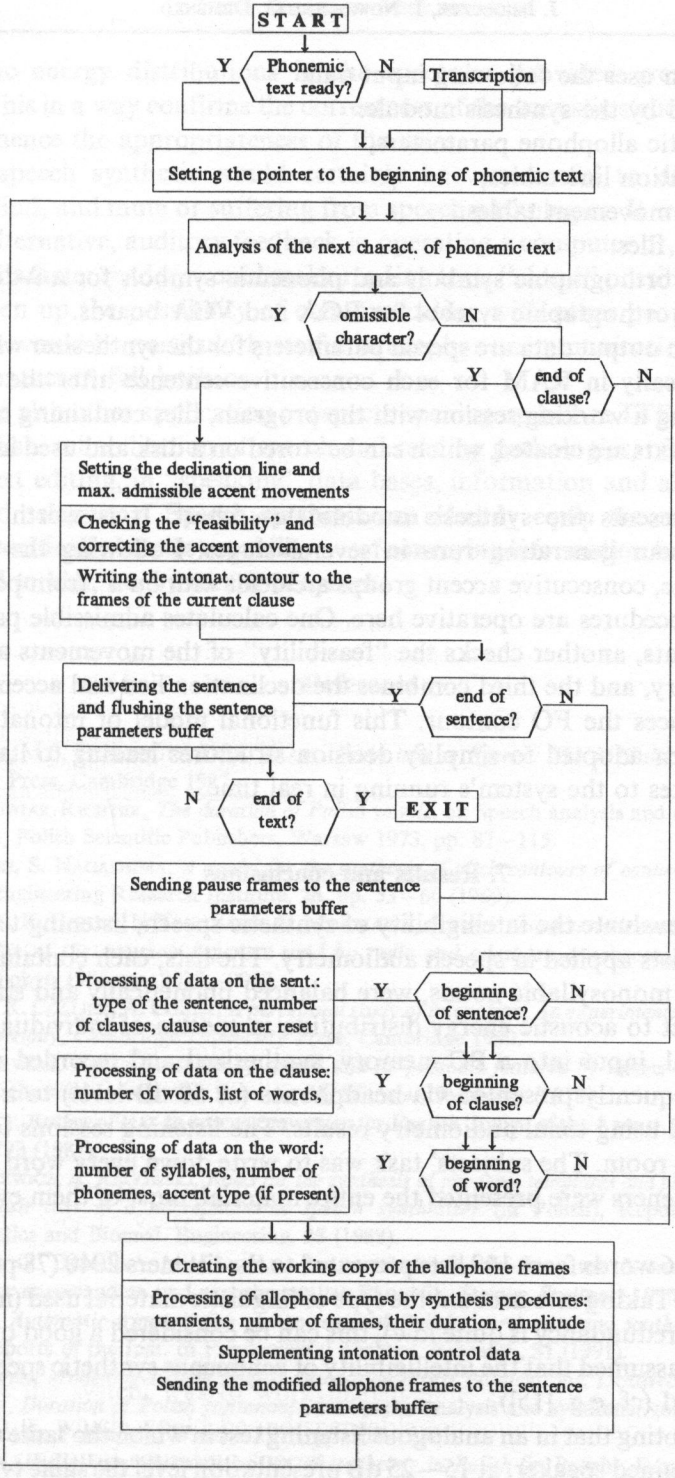
1) Peak position in the stressed vowel: **late**.
2) The rate of accent movements: **fast** rises, **fast** or **slow** falls, depending on the duration of the accent group and its distance from the beginning of the clause.
3) FO range negatively correlated with the number of the accent group and with the number of unaccented syllables occurring between two consecutive accented syllables. The range can never be greater than in the first or last accent group.

The choice of rules referring to the **last** accent group depends on the punctuation mark at the end of the clause. Separate sets of rules have been prepared for declarative sentences (e.g. **early** peak in the penultimate accented syllable with FO slope steepness conditioned by the duration of the final accent group) and for interrogative sentences. In the latter case, depending on question type (whether general or specific), a very fast FO rise (ranging from 50 to 80 Hz) or only a small rise (of the order of 10–20 Hz) have been applied, each preceded by a drop of FO down to a value close to Fmin.

Control parameter values have been verified using data from listening tests and from additional acoustic analyses.

6. Software

The main body of the program and all its modules were written in Turbo Pascal, v. 5.0. The system makes an integrated mini-environment for automatic conversion of Polish orthographic texts into the acoustic speech signal.


START
Phonemic text ready?
Y
N
Transcription
Setting the pointer to the beginning of phonemic text
Analysis of the next charact. of phonemic text
omissible character?
Y
N
end of clause?
Y
N
Setting the declination line and max. admissible accent movements
Checking the "feasibility" and correcting the accent movements
Writing the intonat. contour to the frames of the current clause
Delivering the sentence and flushing the sentence parameters buffer
end of sentence?
Y
N
end of text?
N
Y
EXIT
Sending pause frames to the sentence parameters buffer
Processing of data on the sent.: type of the sentence, number of clauses, clause counter reset
beginning of sentence?
Y
N
Processing of data on the clause: number of words, list of words,
beginning of clause?
Y
N
Processing of data on the word: number of syllables, number of phonemes, accent type (if present)
beginning of word?
Y
N
Creating the working copy of the allophone frames
Processing of allophone frames by synthesis procedures: transients, number of frames, their duration, amplitude
Supplementing intonation control data
Sending the modified allophone frames to the sentence parameters buffer

The program uses the following input data:

1. Files used by the synthesis module:
   a) synthetic allophone parameters,
   b) declination line tables,
   c) accent movement tables.
2. Auxiliary files:
   a) Polish orthographic symbols and phonemic symbols for a 9-dot printer.
   b) Polish orthographic symbol for EGA and VGA boards.

The program output data are speech parameters for the synthesizer which are files created dynamically in RAM for each consecutive sentence after the start of the synthesis. During a working session with the program, files containing orthographic and phonemic texts are created, which can be stored on a disk and used in subsequent sessions.

Figure 1 presents the synthesis module flow chart. It is worth noting that intonation contour generation runs in several stages. Following the selection of a declination line, consecutive accent groups are dealt with on a "from peak to peak" basis. Three procedures are operative here. One calculates admissible parameters of accent movements, another checks the "feasibility" of the movements and modifies them, if necessary, and the third combines the declination line and accent movement data and produces the FO contour. This functional model of intonation contour shaping has been adopted to simplify decision structures leading to its generation, which contributes to the system's running in real time.

7. Results and conclusions

In order to evaluate the intelligibility of synthetic speech, listening tests were run using 10 word lists applied in speech audiometry. The lists, each containing 24 fairly frequent Polish monosyllabic nouns, were balanced phonetically and structurally as well with respect to acoustic energy distribution in spectra of individual phones.

The material, input into a PC memory, synthesized and recorded on magnetic tape, was subsequently presented via headphones (at 50 dB level) to a panel of 36 subjects selected using tonal audiometry results. The listening sessions took place in a sound-treated room. The subjects' task was to write down every word heard. Only a number of listeners were presented the entire material, most of them evaluated 2 to 3 lists.

In all, of 2616 words from 109 lists presented to the listeners 2040 (78 per cent) were heard correctly. Taking into account the type of linguistic material used (monosyllabic words in which redundancy is quite low), this can be considered a good overall result. It can be safely assumed that the intelligibility of **continuous** synthetic speech would be very high indeed (cf. e.g. [15]).

It is worth noting that in an analogous listening test in which the same material was recorded by a trained speaker, at 15–25 dB presentation level **the same** types of errors were most common: **both** in natural and synthetic utterances certain syllabic structures

and acoustic energy distributions in the spectrum provided greater recognition problems. This in a way confirms the correctness of the acoustic structure of synthetic words and hence the appropriateness of the rules.

Text-to-speech synthesis would certainly be very useful to disabled persons, especially blind, and mute or suffering from speech pathologies. It would offer to the former an alternative, auditory feedback in operating a computer or, if coupled to an automatic character reader, would make reading texts "aloud" possible. To the latter, it would open up the possibility of communicating with other people by means of voice, which would be particularly useful in telephone communication where there is no visual contact to fall back on.

There are also other application prospects opening up before TTS synthesis, e.g. in teaching Polish, rehabilitation of speech and reading pathologies, office automation, computer text editing, in "speaking" data bases, information and alarm systems, in banking and telephony. **Speech synthesis can already equip the machine with the up-to-now specifically human capability — of conveying information to man in the form most natural to him, i.e. by means of speech.**

Reference

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

[2] L. FRĄCKOWIAK-RICHTER, *The duration of Polish vowels*, in: Speech analysis and synthesis, vol. 3, ed. W. Jassem, Polish Scientific Publishers, Warsaw 1973, pp. 87–115.

[3] H. FUJISAKI, S. NAGASHIMA, *A model for the synthesis of pitch contours of connected speech*, Annual Bulletin, Engineering Research Institute, **28**, pp. 53–60 (1969).

[4] H. FUJISAKI, K. HIROSE, N. TAKAHASHI, H. MORIKAWA, *Acoustic characteristics and the underlying rules of intonation of the common Japanese used by radio and television announcers*, Proceedings IEEE ICASSP, Tokyo 1986, pp. 2039–2042.

[5] J. t' HART, R. COLLIER, A. COHEN, *A perceptual study of intonation. An experimental-phonetic approach to speech melody*, Cambridge University Press, Cambridge 1990.

[6] G. KIEŁCZEWSKI, *Digital synthesis of speech and its prosodic features by means of a microphonemic method*, Institute of Informatics Reports, **65**, Warsaw University (1978).

[7] D.H. KLATT, *Review of text-to-speech conversion for English*, Journ. of the Acoust. Soc. of America, **82**, pp. 737–793 (1987).

[8] K. ŁUKASZEWICZ, A. RĘGOWSKI, *Rules for the synthesis of phoneme templates and phonetic transcription of the Polish text in a microphonemic speech synthesizer* (in Polish), Reports of the Inst. of Biocybernetics and Biomed. Engineering, **25** (1988).

[9] B. MÖBIUS, G. DEMENKO, M. PÄTZOLD, *Parametrische Beschreibung von Intonations Konturen*, in: Beitrage zur angewandten und experimentallen Phonetik, Steiner, Stuttgart 1990, pp. 109–125.

[10] I. NOWAK, *Automatic transcription of non-regional Polish (north-east and south-west varieties)*, (in Polish), Reports of the Inst. of Fundamental Technol. Research, **31** (1991).

[11] J.R. de PIJPER, *Modelling British English Intonation*, Foris Publications, Dordrecht 1983.

[12] L. RICHTER, *Duration of Polish consonants*, in: Speech analysis and synthesis, vol. 4, ed. W. Jassem, Polish Scientific Publishers, Warsaw 1976, pp. 219–238.

[13] L. RICHTER, *Preliminary characterization of isochrony in Polish* (in Polish), Reports of the Inst. of Fundamental Technol. Research, 4/1983.

[14] L. RICHTER, *Statistical analysis of the rhythmical structure of utterances in Polish speech* (in Polish), Reports of the Inst. of Fundamental Technol. Research, 8/1984.

[15] N. SCHIAVETTI, R.W. SITLER, D.E. METZ and R.A. HOUDE, *Prediction of Contextual Speech Intelligibility from isolated word intelligibility measures*, Journal of Speech and Hearing Research, vol. 27, Number 4, Dec. 1984, pp. 623–626.

[16] M. STEFFEN-BATOGOWA, *Automation of phonemic transcription of Polish texts* (in Polish), Polish Scientific Publishers, Warsaw 1975.

[17] N. THORSEN, *Stress group patterns, sentence accents and sentence intonation in Southern Jutland — with a view to German*, Annual Report of the Institute of Phonetics, University of Copenhagen, 23, pp. 1–85 (1989).