PROSODIC ASPECTS OF POLISH WORD SYNTHESIS

JANUSZ IMIOŁCZYK RYSZARD CIARKOWSKI

Acoustic Phonetics Research Unit, Institute of Fundamental Technological Research, Polish Academy of Sciences, Noskowskiego 10, 61-704 Poznań, Poland

ABSTRACT

Four Polish words with varying number of syllables (dał, dobra, normalny and naturalnie) were synthesized using a COMPUTALKER CT-1 speech synthesizer. For each of the words 8 basic Polish intonation patterns were obtained by appropriately controlling the F0 parameter. Three variants of the intonation patterns were prepared (a quasi-natural variant and two types of approximation) of which the quasi-natural variant, elaborated on the basis of F0 values extracted from natural utterances of the four words, served as the model for the remaining two. The total of 70 synthetic intonation patterns were tested for recognizability and naturalness in a listening experiment. On this basis, the optimum approximation type was determined and the contours most typical for Polish questions and statements were selected for further research in the synthesis of intonation.

INTRODUCTION

The perceptual impression of accent is generally claimed to result from variations in fundamental frequency, duration and intensity within the vowel segment of the accented syllable ([2], [6], [10]). Of these three parameters, F0 has been found to play the most important role in signalling accent ([7], [9], [10], [14]), intensity having the least significant effect ([2], [9], [10]).

In analyses of linguistic functions of intonation two most general types of utterances are distinguished: (1) unfinished (general interrogative, continuative) and (2) finished (statements, demands, specific questions). Utterances of the first type are usually characterized by a rising intonation and utterances of the second type – by a falling intonation. In perceptual identification of each of these types the following three factors are of particular importance:

1) F0 level at the turning point, i. e. the point immediately preceding an F0 rise or fall
2) direction of F0 change and
3) its range.

For example, subjective impression of a general question is the stronger, the greater the F0 increase within the accented syllable ([8], [11]) and the higher the F0 value at the turning point ([8]).The effect of the F0 value at the turning point may even be more relevant than that of the final F0 rise, especially if the latter's range is relatively small. Moreover, the perceptual impression of accent is the stronger, the faster the F0 rise within the accented syllable ([13]).

TECHNICAL BASES OF WORD SYNTHESIS

For the purposes of the present experiment, four Polish words (dał, dobra, normalny and naturalnie) were synthesized using a COMPUTALKER CT-1 formant speech synthesizer controlled – via a minicomputer MERA 303 configuration (Fig 1; cf. also [4]) – by specially developed software. COMPUTALKER CT-1 simulates the transfer function of the vocal tract by means of formant filters connected in series. Apart from the synthesizing unit, composed of noise and glottal tone generators as well as nasal, noise and formant filters, it also includes a logistics unit the function of which consists in:

a) conversion of 8-bit digital parameters into the analog form controlling the elements of the synthesizing unit
b) short-term memory storing of the parameters.

A set of 9 parameters (amplitudes of the glottal tone, aspiration noise and nasal resonance, F0, F1, F2, F3, amplitude and frequency of the friction noise) has to be input to the synthesizer in digital form (1 byte per 1 parameter) for the synthesizer to be controlled. This is done by means of a parallel bus of data (each control parameter is given a 4-bit code) termed "frame". The rate of control data transmission in the system corresponds to an average speech rate (1 frame per 10 ms).

WORD SYNTHESIS. PRINCIPLES CONCERNING DURATION AND AMPLITUDE OF THE GLOTTAL TONE

Word synthesis was carried out using the existing library of synthetic diads of the

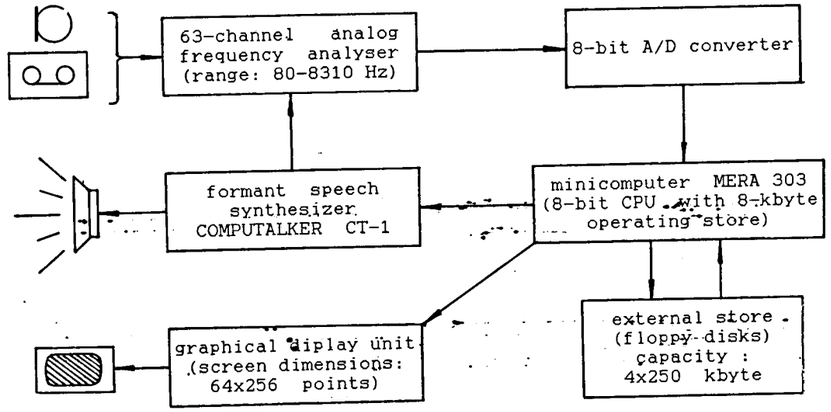


Fig. 1. Minicomputer configuration for speech synthesis and analysis

CV and VC type ([3]). Appropriate transient segments were inserted between the diads to form natural-sounding words. At the initial stage each of the synthetic words received "intensity-durational" stress only, with constant F0 at 120 Hz over their whole length. Amplitude changes in the formant tract were of segmental character in almost all cases: within the steady state of any monosegmental phone amplitude values were not varied. Since the role of intensity in signalling accent is marginal (cf. above), amplitude values within stressed vowels were also held flat (the same approach was adopted by Abramson [1] and Mattingly [12]).

Durations of individual phones making up the four words were determined on the basis of the results given by Richter [15]. Preliminary listening tests led to the following duration of the steady-state part of the stress-bearing vowel within the total word duration:

dał	– 190 ms (total	– 480 ms)
dobra	– 130 ms (total	– 670 ms)
normalny	– 120 ms (total	– 910 ms)
naturalnie	– 90 ms (total	– 1060 ms)

SYNTHESIS OF INTONATION

For the purposes of intonation synthesis, 8 contours (corresponding to the most typical Polish word intonations) were selected from the inventory of Polish intonemes put forward by Steffen-Batogowa ([16]). These were:

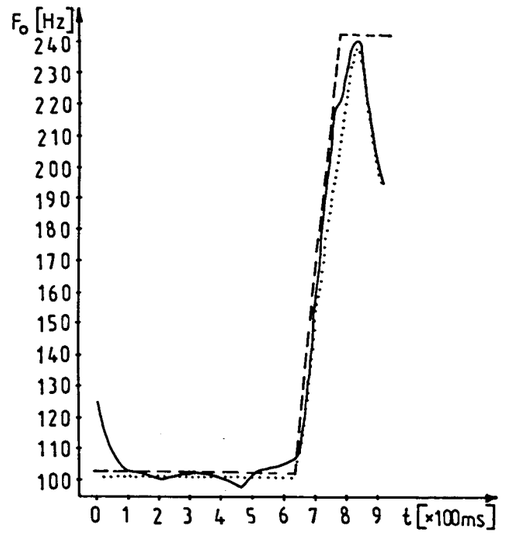
1) low rising (LR)
2) full rising (FR)
3) high rising (HR)
4) level (L)
5) full falling (FF)
6) low falling (LF)
7) full rising-falling (FRF)
8) low rising-falling (LRF)

Natural utterances of the four words, spoken with all these intonations by a skilled phonetician, were tape-recorded. As the utterances of the words dał, normalny and naturalnie containing the composite contours (LRF and HRF) sounded somewhat artificial, they were excluded from further examination. F0 patterns in the remaining utterances were analysed using a TM3 pitch meter. The resulting sequences of absolute F0 values, each corresponding to a 10 ms interval, were used to synthesize quasi-natural intonations which served as the model for the following two types of approximation of natural F0 contours:

1) Approximation by a broken line (A1) consisting of three or four segments. This is the most frequently applied approximation (cf. e.g. [5], [11]). Four-segment broken line was utilized in rising-falling intonations. Level intonations were approximated by a straight line.
2) Step-wise approximation (A2), in which a rapid change in F0 (10 Hz/10 ms), carried out within the accented syllable, occurred between two level segments.

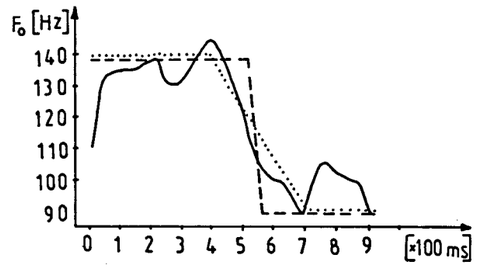
Irrespective of F0 pattern type, neither of the two parameters responsible for signalling accent (i.e. duration and amplitude) was modified. It was assumed that variations in length of the accented vowel occurring among different types of F0 patterns do not affect accent perception and intonation type identifiability in any significant way.

Altogether, 70 synthetic intonation patterns were prepared. The total fundamental frequency range utilized in the synthesis covered frequencies from 77 Hz to 250 Hz. In the majority of cases, however, F0 values were not lower than 88 Hz and did not exceed 240 Hz.



quasi-natural contour ——————
approximation by a broken line ·····
step-wise approximation – – – – –

Fig. 2. NORMALNY - full rising intonation



quasi-natural contour ——————
approximation by a broken line ·····
step-wise approximation – – – – –

Fig. 3. NORMALNY - low falling intonation

Examples of synthetic F0 contours are given in Figs. 2 and 3.

PERCEPTUAL EVALUATION OF SYNTHETIC INTONATIONS

Recognizability and naturalness of all the synthetic intonations were evaluated perceptually, in two listening tests, by 18 subjects divided into two panels (A and B). The task of panel A, consisting of 8 persons with professional experience with respect to speech melody, was to identify the type of intonation pattern presented (LR, FF. L etc.). Panel B, composed of 10 "naive" subjects, was to qualify the synthetic utterances as questions or statements.

Results

Panel A. For the total of 560 identification trials (70 patterns x 8 subjects) 401 correct responses were obtained (ab. 72 %). Identification rate of the intonations in the four words was the highest with the step-wise approximation (76 %) The quasi-natural intonations and the contours approximated by a broken line were recognized correctly in 69 % and 71 %, respectively.
Of the 8 types of synthetic intonations, low rises were recognized most efficiently (92 per cent of correct responses), whereas the composite patterns yielded the poorest identification results (only 21 per cent of correct responses were obtained for full rise-falls). The recognition scores for the remaining intonations were as follows: low fall - 78 %, level - 69 %, full fall - 81 %, high rise - 66 %, full rise - 73 %, and low rise-fall - 33 %.

Panel B. Two of the three synthetic rising intonations (full rise and high rise) proved to be nearly equally effective in signalling a question: they were perceived as interrogative 98 % of the time. With the exception of the low rise, which was misidentified most frequently, all the remaining intonations were commonly judged (from 88 to 100 per cent of responses) as typical for statements.
As the case of the level intonation indicated, an F0 fall is not an indispensable condition for an utterance to be perceived as statement-like.

DISCUSSION

In 87 % of the erroneous responses, the direction of F0 change was recognized correctly and the error pertained only to the range of variation . With the low intonations, however, the contrary tendency was observed to occur. Even though the LR and LF patterns were apparently fairly easy to identify (92 and 78 per cent of correct responses, respectively), they were also quite often confused with the level pattern. The factor responsible for this perceptual similarity was most probably the relatively small range of F0 variation in LR and LF.
As stated above, identification scores obtained for the complex patterns (LRF and FRF) were the lowest of all. Due to their less frequent occurrence in Polish one-word utterances, the two were often perceived as simple falling intonations.
The reason why the step-wise approximation provided the best identification results of intonation types was the characteristic, abrupt change in F0 preceded and followed by relatively long (hence easily perceptible) segments within which F0 value remained constant (level). The drawback of the intonation patterns thus produced was the peculiar "singing effect" of the utterances containing them.
Responses given by panel B pointed to the occurrence of a tendency for preferring the "statement" alternative. An intonation rise within the accented syllable was found to be the necessary condition of the "question" response. Moreover, the rise had to be characterized by a sufficiently wide range or a sufficiently high F0 level within the pre-accentual segment of the word. At least one of these conditions was met with FR and HR intonations, which were almost unanimously judged as typical for questions. On the other hand, a considerable divergence of responses occurred with LR intonations which, owing to the low F0 level within the pre-accentual segment and the small F0 increase, were perceived as indicating statements by the majority of subjects (cf. [11]).

CONCLUSIONS

The results obtained in the present experiment provide a number of cues which are essential for further research in Polish intonation synthesis. They suggest, among others, that due to both relatively high recognition rate and naturalness, the optimum approximation variant is the one utilizing approximation by a broken line. Of the two rising intonations judged as typical for Polish general interrogative utterances, the FR should be selected as the model one, as it is characterized by a wider frequency of usage and, thus, is more neutral. For similar reasons, the LF (and, perhaps, the FF) should be chosen as the model "declarative" intonation(s).

REFERENCES

[1] A.S. Abramson, Static and dynamic acoustic cues in distinctive tones, Lang. & Speech 21, 1978, 319-325.
[2] C. Adams, R.R. Munro, In search of the acoustic correlates of stress: Fundamental frequency, amplitude and duration in the connected utterances of some native and non-native speakers of English, Phonetica 35 1978, 125-156.
[3] R. Ciarkowski Minicomputer MERA 303-controlled synthesis of selected Polish diads and their perception (in Polish), IFTR Reports 7/1984, Warsaw.
[4] R. Ciarkowski, J. Imiołczyk, Analysis-aided formant speech synthesis, MELECON '85, vol.II, 171-173, North Holland, 1985.
[5] J.E. Clark, A low-level speech synthesis by rule system, J. of Phonetics 9, 1981, 451-476.
[6] P. Denes, J. Milton-Williams, Further studies in intonation, Lang. & Speech 5, 1962, 1-14.
[7] L. Dukiewicz, Intonation of Polish Utterances (in Polish), Ossolineum, Wrocław, 1978.
[8] K. Hadding-Koch, M. Studdert-Kennedy, An experimental study of some intonation contours, Phonetica 11, 1964, 175-185.
[9] W. Jassem, J. Morton, M. Steffen-Batog, The perception of stress in synthetic speech-like stimuli by Polish listeners, [in:] Speech Analysis and Synthesis (W. Jassem, ed.), vol.1, 289-308, Warsaw, 1968.
[10] I.Lehiste, Suprasegmentals, The M.I.T. Press, Cambridge, Massachusetts, 1970.
[11] W. Majewski, R. Blasdell, Influence of fundamental frequency cues on the perception of some synthetic intonation contours, J. of Acoust. Soc. of Amer. 45, 1969, 450-457.
[12] I.G. Mattingly, Synthesis by rule of prosodic features. Lang. & Speech 9, 1966, 1-13.
[13] S. Öhman, Word and sentence intonation: A quantative model, Speech Transmission Laboratory QPSR 2-3, 1967, 20-54.
[14] J.P. Olive, Fundamental frequency rules for the synthesis of simple declarative English sentences, J. of Acoust. Soc. of Amer. 57, 1975, 476-482.
[15] L. Richter, Statistical analysis of the rhythmical structure of utterances in the Polish speech (in Polish), IFTR Reports 7/1984, Warsaw.
[16] M. Steffen-Batogowa, Versuch einer strukturellen Analyse der polnischen Aussagemelodie, Zeitschr. f. Phon. u. allgem. Sprachwiss. 19, 1966, 398-440.