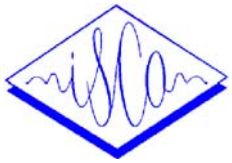


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

MODELLING INTONATIONAL PHRASE STRUCTURE WITH ARTIFICIAL NEURAL NETWORKS

Grażyna Demenko, Wiktor Jassem^*
*A. Mickiewicz University in Poznań, Institute of Linguistics, Poland
lin@main.amu.edu.pl
^Polish Academy of Sciences, Institute of Fundamental Technological Research, Poznań
wjassem@main.amu.edu.pl

10.21437/Eurospeech.1999-172

ABSTRACT

A model of intonation for Polish has been created on the basis of a general theory of suprasegmentals and on experiments using isolated utterances as well as continuous speech. An intonational phrase consists of an optional prenuclear tune and an obligatory nuclear tune. A training of a three-layer MLP network was performed distinguishing 9 nuclear accents: HL, ML, LL, HM, LH, LM, MH, MM, LHL and 2 secondary prenuclear accents: High (H) and Low (L). A total of 1600 structures (in constructed phrases) were used for training, and 430 for verification. The average score for training and testing was 82 percent. In continuous speech the following structures were postulated: H and L for prenuclear intonation and for nuclear intonations: R (rising), F (falling), MM (level), LHL (rising-falling). For the testing set, a score between 79 and 83 per cent was obtained. In both classifications, an 11-element vector was used to describe the intonational structures under analysis.

1. INTRODUCTION

Polish accent is realized by intonation ([1]). There are, as probably in every intonation language, a finite number of melodic patterns, each pattern forming an intonation phrase ([2], [3], [5]). The phrase includes, in Polish, exactly one nuclear tune, which is in final position. It may be preceded by one or more strong prenuclear tunes. The first, or only syllable of the nuclear tune and the strong prenuclear tune bears pitch accent. The prenuclear tune may begin with an phrase-initial accented fragment, or an unaccented fragment followed by an accented one. The accent in the prenuclear tune is regarded as secondary, whilst the accent in the nuclear tune is primary. The nuclear tune begins with the last tonally accented syllable in the phrase. A further accent, if any, is durational rather than tonal.

The present paper proposes a parametrization of the time course of fundamental frequency based on a model which has been verified in perceptual experiments and an analysis of F_0 traces using Polish speech material.

2. THE EXPERIMENTAL MATERIAL

The materials analyzed in the present paper are of two kinds: (a) constructed phrases, and (b) read texts. The constructed phrases involving 2 prenuclear and 9 nuclear tunes (40 in all) were first spoken by a phonetician, then imitated (twice) by one panel of 26 students and judged for linguistic equivalence to the model by a different panel of 20 students.

In order to test the linguistic equivalence of the 2080 imitations, the perceptual experiment consisted in rating the imitations on a 5-step scale of similarity. The imitations that were not in agreement with the model (6%) were discarded. The continuous speech in this experiment included fragments from 2 general-interest press excepts lasting 9 minutes each, and read by 3 trained speakers.

The listening panel included two groups: (1) 25 students, who were naive listeners without phonetic or linguistic training, and (2) 5 subjects with phonetic training. The experiment was performed in two stages.

First, accented syllables were marked, and then, separately, borders between phrases. The listener was presented with the text of the reading, all printed with lower-case letters and with no punctuation marks. The listener's task consisted solely, first, in indicating the syllable he/she judged as accented, and second, in indicating the border between fragments of the text that he/she considered to be internally coherent and relatively distinct from the neighbouring parts of the text. Additionally, the phonetically trained listeners indicated the final fragment of each phrase beginning with the last accented syllable as having one of the following types of intonation: rising, falling, level or rising-falling. One of the phoneticians additionally indicated secondary accents.

The classes of nuclear tunes obtained in the preceding two parts of the materials were correspondingly conflated: LH, MH and LH formed one class of rising tunes, R, and HL, HM, ML formed the class of falling tunes, F.

The overall results of the experiment included the following types of accent: H (496 cases), L (35 cases), R (209 cases), F (175 cases), MM (34 cases), LHL (9 cases) and anacrusis P (78 cases). R, F, MM and LHL open nuclear tones.

The experiment testified a high degree of uniformity among listeners as regards the segmentation of a speech signal.

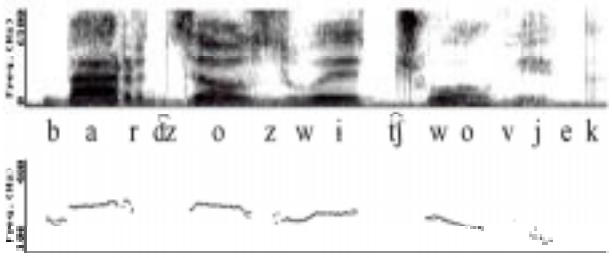
A comparison of the contours shows that the H prenuclear tune begins near a local (or global) maximum, while an L prenuclear tune begins near a local (or global) minimum in the F_0 trace. Extreme values of the F_0 parameter on nuclear accents usually occurred near the global extremes of the F_0 pattern.

Fig. 1a represents a phrase with nuclear tune of type ML (on the syllable *czło* and two prenuclear H accents (syllables *bar* and *zły*).

Fig. 1b shows a phrase with nuclear tune of type HL (on the syllable *znak*) and two L-type prenuclear accents (on *bar* and *do*).

Prenuclear syllables have little F_0 variation, which is distinctly stronger in the nuclear syllables (between about 12 and several tens of Hz), except for the level MM nuclear tune. An acoustic analysis permitted the postulation of distinctive parameters which distinguish nuclear and prenuclear accents. The results were subjected to statistical Analysis of Variance and Discriminant Analysis.

a)



b)

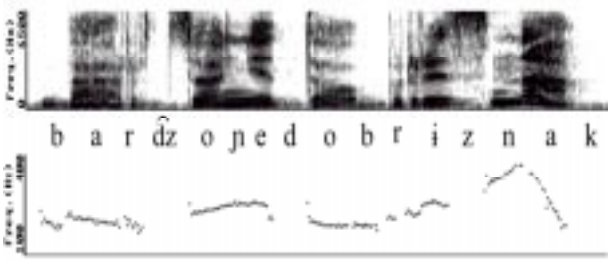


Fig. 1a represents a phrase with nuclear tune of type ML (on the syllable czło and two prenuclear H accents (syllables bar and zły).

Fig. 1b shows a phrase with nuclear tune of type HL (on the syllable znak) and two L-type prenuclear accents (on bar and do).

In most imitations of these two phrases, there were no more than two prenuclear tunes. An L prenuclear tune was occasionally replaced by H in an imitation (but not vice versa). This probably indicates that the prenuclear tunes are less categorical than the nuclear.

3. MODELLING INTONATIONAL PHRASE

In order to model temporal pitch variation in the phrase, it is necessary to adopt a number of premises which relate to perceptual-acoustic features of the signal.

1. The basis for a classification/recognition of pitch variations is a complete recurrent pitch pattern. Analysing the pitch variations syllable by syllable does not result in acceptable classification.

The following 12 classes have been assumed:

a) constructed phrases: nine nuclear tunes, two prenuclear tunes, H and L, and a class P, which is one or more unaccented syllable preceding a prenuclear tune,

b) for the read text: LH, MH and LH formed one class of rising tunes, R, and HL, HM, ML formed the class of falling tunes, F; MM, LHL, two prenuclear tunes, H and L, and a class P.

2. The acoustic features of pitch patterns are essentially defined by variations of F_0 on vowels and relations between the syllabic pitches.

3. In continuous speech, finding phrase boundaries is facilitated by such acoustic features as elongation of syllables and reduction of energy level.

The vector of structural features (extracted from the original measurements) which contains information about the suprasegmental structure within a vowel/syllable or a sequence of syllables can form the basis for the parametrization of the pitch contour.

4. PARAMETRIZATION OF PITCH CURVES

As in other, more ambitious systems of language processing (such as, e.g., Verbmobil, [6]), intonation is described by a vector of structural features based on measurements of F_0 traces. Rather than postulating a possible, extremely complex combination of several tens or even hundreds of features (such as the relations of initial to final pitch in vowels and/or syllables), the structures under analysis here are described in terms of a vector of a few elements. An eleven-element feature vector was used, with the features defined below.

The first two features determine the direction of pitch change: falling, rising or rising-falling.

1. $x_1 = V_p - F_e$

x_1 describes the difference between the initial value of F_0 (V_p) on the first vowel of the tune and the value of F_0 at the extremum (F_e).

2. $x_2 = F_e - F_k$

x_2 is the difference between the the F_0 value at the extremum, F_e, and at the final point of the tune, F_k.

3. $x_3 = F_{max} - F_{min}$

x_3 determines the total range of variation between the maximum value of F_0 and its minimum value within the tune.

x_3 is used to distinguish tunes with a large range, such as LH from those with a small range, such as LM.

4. $x_4 = F_{sr} - F_{srg}$

x_4 refers to the difference between the mean F_0 value in the tune and a global mean value for the phrase, and is used to distinguish prenuclear H from prenuclear L.

5. $x_5 = F_{min} - F_{ming}$

This expression determines the difference between the minimum (F_{min}) in the tune and the mean minimum F_0 for the given voice (F_{ming}).

Fig. 2a presents, by way of an example, the data describing 9 nuclear accents in the design direction DF ($x_2 = F_e - F_k$) and minimal fundamental frequency FMIN ($x_5 = F_{min} - F_{ming}$). The direction coordinate separates rising from falling intonations, e.g. LH from HL. F_{min} separates low from high intonations, e.g., LM from MH. The data for the 9 nuclear tunes in Fig. 2b describe the 9 nuclear accents in the design extremum ($x_1 = V_p - F_e$) and range ($x_3 = F_{max} - F_{min}$). The extremum coordinate (EKS) separates falling from rising-falling intonations, e.g. HL from LHL, the range coordinate (ZAK) separates wide from narrow intonations, e.g. LM from LH. As can be gathered from a visual inspection of the dispersions of the individual classes, the sets are not linearly disjoint. There are two global values, F_{ming} and F_{srg}, i.e., the average minimum value and the overall mean for the given voice.

a)

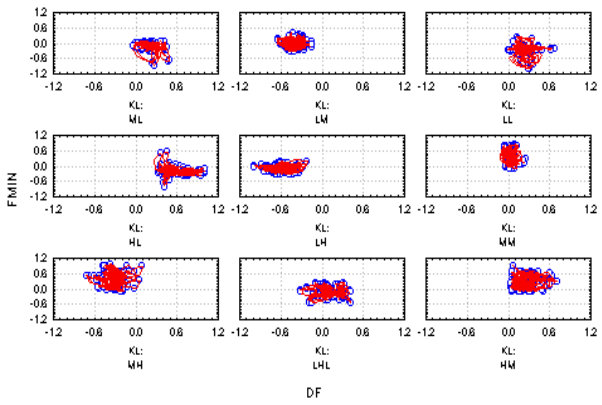


Fig. 2a. Data describing the 9 nuclear accents in the co-ordinate system: direction DF ($x_2 = F_e - F_k$) and FMIN ($x_5 = F_{min} - F_{ming}$).

b)

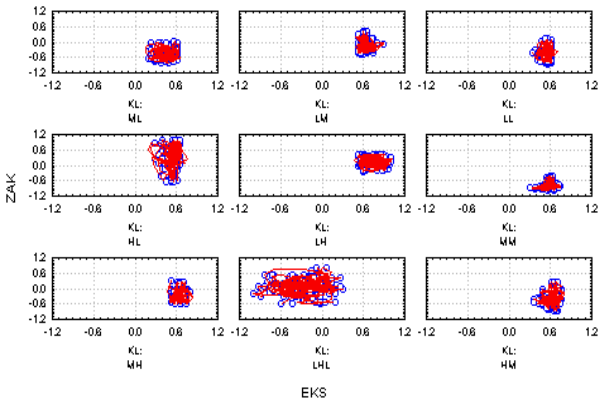


Fig. 2b. Data describing the 9 nuclear accents in the co-ordinate system: ekstremum EKS ($x_1 = V_p - F_e$) and range ($x_3 = F_{max} - F_{min}$).

Features 6, 7 and 8 are used for distinguishing nuclear tunes.

6. $x_6 = V_{pe} - F_{ke}$

x_6 is the variation of F_0 within the syllable with maximum F_0. In rising nuclear intonations, this maximum appears near the end of the phrase.

7. $x_7 = |V_p - F_k| - |V_{pa} - F_{ka}|$

x_7 is the difference between the global variation range in the tune $|V_p - F_k|$ and the variation range within the accented syllable $|V_{pa} - F_{ka}|$.

8. $x_8 = |V_{pa} - F_{ka}| - |V_{ka} - F_{kr}|$

x_8 is the difference between the variation of F_0 within the accented syllable $|V_{pa} - F_{ka}|$ and the variation $|V_{ka} - F_{kr}|$ measured from the end of the accented syllable to the end of the phrase.

The features 9, 10 and 11 are related to duration and energy of the last vowel in the phrase.

In part 1 of the material, all the utterances were spoken separately as complete intonation phrases. In continuous speech, borders between intonation phrases are not always marked by a silence, though location of the border may be indicated by temporal expansion of the final syllable in a phrase and a decrease of energy in the final fragment.

9. $x_9 = t_i$

x_9 is the normalized duration of the last vowel in the phrase related to the mean and the standard deviation of duration for the vowels in the tune.

The phrase-final vowel has increased duration. Extra duration may also be expected, generally, in vowels in nuclear tunes, i.e., the phrase-final tunes, contrasting with shorter durations in preictic tunes.

Fig. 3 shows normalized duration of the final vowel in individual classes of tunes.

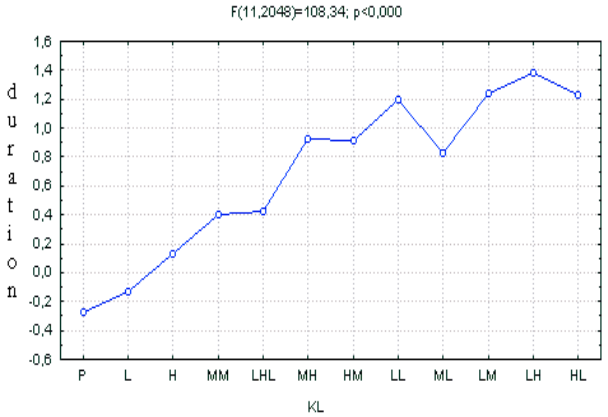


Fig. 3. Normalized duration of the final vowel in individual classes of tunes.

10. $x_{10} = DF/DT$

x_{10} relates the slope of F_0 within the last vowel, where $DF = V_p - F_a$ and DT is the vowel's duration.

11. $x_{11} = E_i$

x_{11} is the normalized energy in the last vowel of the tune being classified.

5. CLASSIFICATION

Several types of neural networks were used for the classification of the nuclear accent: (a) probabilistic, (b) MLP with radial activation functions (90 hidden neurons) (c) classical four-layer network with 6 hidden neurons in the first layer and 6 in the second layer, (d) classical three-layer MLP ([4]).

A detailed training of a three-layer MLP network was performed using a back-propagation method, distinguishing 9 nuclear accents and 2 secondary prenuclear accents. A total of 1600 structures were used for training, and 430 for verification. With 7 neurons the global squared error of classification was 0,11 in training and 0,13 in verification. With 11 neurons, the error in training was 0,08, and the error in verification 0,09. A further increase in the number of neurons gave no improvement (Fig. 4). The average score for training and testing was 82 percent. Initial unaccented syllables were the most difficult to classify (67 percent score in training and 60 percent in verification). Primary accents of the types LH, HL and LHL were classified satisfactorily (over 80 percent score). The main problem in the classification in continuous speech is to distinguish secondary from primary accents. Best results were obtained for a network with 9 hidden neurons. For the testing set, a score between 79 and 83 per cent was obtained. In both classifications, an 11-element vector was used to describe the intonational structures under analysis.

Most of the prenuclear tunes were classified as H, and the most frequent nuclear tune was a fall. Those of our acoustic features which are directly related to the end of the phrase do not have a very strong predictive power, which was the main source of some errors in the differentiation between nuclear and prenuclear tunes in the read texts.

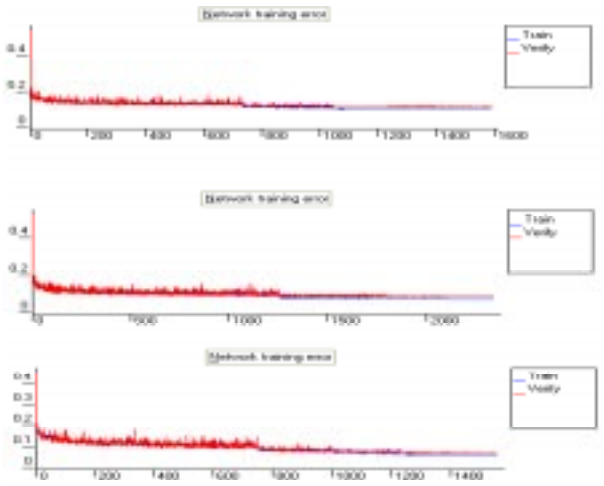


Fig. 4. Error plot in MLP network training.
a) 7 neurons (E_u = 0,11, E_t = 0,13)
b) 11 neurons (E_u = 0,08, E_t = 0,09)
c) 16 neurons (E_u = 0,069, E_t = 0,08)

6. CONCLUSIONS

The acoustic analyses and the classifications of the tunes confirmed the apriori structure based on general premises which posits a sequential model of the intonational phrase for Polish.

Rather high scores of the classification procedure confirmed the bases of parametrization into some twelve features and justified its application both to isolated utterances and continuous texts.

Future research should extend the present study by including spontaneous speech.

ACKNOWLEDGMENTS

The theoretical and the experimental part of this work was funded by contract EP-20288 CRIT2. The testing procedure using neural networks was funded by contract 8T11E 042 KBN.

REFERENCES

[1] Demenko G., Jassem W., Krzyśko M. (1988) Classification of basic F_0 patterns using discriminant functions, Phonetica 41, 1-12.

[2] Hirst D., DiCristo A. (1996), Intonation Systems, Cambridge Univ. Press, Cambridge.

[3] Jassem, W. (1996) A quantitative analysis of Standard British English Nuclear Tunes, J. of Quantitative Linguistics vol. 3, 229-243.

[4] Morgan D. P., Scofield Ch., (1992) Neural Networks and Speech Processing, Kluwer Academic Publishers, Boston/Dordrecht/London.

[5] O'Connor, J. D. and Arnold, F. F. (1973) Intonation of Colloquial English, Longman, London.

[6] Sagisaka Y., Campbell N., Higuchi N. (1997) Computing Prosody, Computional Models for processing Spontaneous Speech, Springer - Verlag, New York.