# A CLASSIFICATION OF POLISH PITCH ACCENTS USING NEURAL NETWORKS.

**Grażyna Demenko*, Wiktor Jassem^**

**A. Mickiewicz University in Poznań, Institute of Linguistics, Poland*

*^Polish Academy of Sciences, Institute of Fundamental Technological Research, Poznań*

**ABSTRACT**

Neural networks were used to classify nuclear and prenuclear tunes. Experiment (1) included imitations of one- or two-syllable phrases representing 9 different nuclear tunes defined in terms of Low, Mid or High tone. Exp. (2) included imitations of longer phrases, some with an unaccented PTs, H or L PTs and the 9 NTs of exp. (1).

All imitations were judged for equivalence with the models, and wrong imitations were discarded. In (1) and (2) one part of the materials was used for training and another part for testing. Exp. (3) included a read text, in which some of the classes of tunes distinguished in exp. (1) and (2) were conflated so as to produce 4 classes of nuclear tunes. The accuracy of classification in the training set and the testing set was similar: The scores were mostly contained within a bracket of approx. 80 to 90 percent.

## 1. THE EXPERIMENTS

The materials analyzed in the present paper are of two kinds: (a) constructed phrases, and (b) read texts. The first type of data was obtained in experiments involving imitations of two kinds of intonation phrases: (1) short, single-word one- or two-syllable utterances involving only nuclear tunes, and (2) longer phrases with prenuclear tunes and/or long nuclear tunes.

The constructed phrases (60 in all) were first spoken by a phonetician, then imitated by one panel of 24 students and judged for linguistic equivalence to the model by a different panel of 20 students. The imitators were chosen at random, including 10 male and 14 female voices.

In order to test the linguistic equivalence of the 1440 imitations, the perceptual experiment, which was extended over a period of one week, consisted in rating the imitations on a 5-step scale of similarity between 0 and 4 points. The grand mean rating was 3.2, and it was noted that longer phrases including prenuclear tunes were somewhat more difficult to imitate.

All imitations that received a rating of better than 2.0 were assumed to agree with the model. Their proportion was 94 %, which can be interpreted as indicating ease in imitating melodic structures in speech.

The imitations that were not in agreement with the model were discarded.

## 2. THE GENERAL BASES OF CLASSIFICATION

An initial inspection of the acoustic data in the form of $F_0$ traces lead to the conclusion that any classification of the utterances with respect to intonation and pitch accent must be holistic in the sense of comprising the entire intonation phrase.

$F_0$ courses in individual syllables and relations between the pitches of neighbouring syllables alone do not lead to a linguistically relevant classification. Such temporal variations of $F_0$ must be included in the descriptions of complete phrases, supplemented by information as to the temporal alignment of certain $F_0$ features with vowels as syllabic cores.

In part 1 and part 2 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.

## 3. NUCLERAR TUNES ALONE

The models and the imitations consisted of simple one- or two-syllable utterances such as *znak* (a sign); *znów* (again); *tak* (yes); *proszę* (please) etc. 9 types of nuclear tune were distinguished, as indicated below. Each type was represented by 210 replications, 120 monosyllabic and 90 disyllabic.

As noted in other languages (Hirst and DiChristo 1998), the assignment of an $F_0$ trace to a class, such as a rise or a fall, etc. is not trivial. Apart from segmental (microprosodic) interactions, the alignment with syllabic cores may be crucial. Also, what are assumed as representing level tunes may show some slight rise or slight fall.

The nuclear tunes were described in terms of 5 features chosen on the basis of the authors' previous experience and as found in the literature (Demenko et al.1988). Three of these features are related to the shape and the range of $F_0$ trace, and the remaining two to the speaker's range and minimum of $F_0$.

A MLP-type three-layer Neural Network with sigmoidal basic elements was used for the training, with back-propagation and momentum element (Morgan et al. 1992). The successive layers were each connected with each. The number of input neurons is defined by the dimension of the input vector (5 features describing the intonation structures). The number of the output neurons is equal, in the classifier, to the number of classes.

Using L for low, M for mid and H for high, the nine nuclear tunes distinguished here were HL (i.e. High-to-Low, etc), ML, HM, MH, LH, LM, MM, LHL and LxL (from L down). This classification is largely based on descriptions of English intonation as proposed by the British tradition of Intonation analysis (see, e.g., O'Connor and Arnoldt 1973, Jassem 1996).

For each class, there is, in theory, one output vector in which only one element is different form the others, but in practice they may even all be slightly different, but one is very distinctly different. Threshold values were therefore used, viz. 0.95 for accept and 0.25 for reject.

The apriori classification was based on results of the perceptual test described above. Several training cycles were performed, with two parameters, viz. h1 as learning rate and h2 as the momentum term.

The following coefficients were selected: for h1, 0.9, 0.6, 0.3, and 0.09, and for h2, 0.6, 0.4. 0.2 and 0.06. A total of 1000 presentations were included in the training set.

The mean classification rate was 86 percent in the training and 83 percent in the test. The nuclear types LH and HL (being acoustically the most distinct) received he best scores.

## 4. NUCLEAR TUNES PRECEDED BY PRENUCLER TUNES

The kind of problem encountered with the classification of prenuclear tunes is illustrated in Fig. 1a and 1b, 2a and 2b.

The phrase is (1) *To był* (H)*całkiem* (H)*niezły i ucz*(H)*ciwy* (ML)*człowiek.* (He was quite a good and honest man) with three H prenuclear tunes and and ML nuclear tune, and (lexically) the same phrase with three L prenuclear tunes: (2) *To był* (L)*całkiem* (L)*niezły i ucz*(L)*ciwy* (ML)*człowiek.*

Each of the two structures is shown twice representing (a) the model and (b) the best imitation. A comparison of the contours shows that the H prenuclear tune begins near a local maximum, while an L prenuclear tune begins near a local minimum in the $F_0$ trace. 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.

The classification in longer phrases with both nuclear and prenuclear tunes is a much more complex problem than the classification in simple phrases with just one accent.

The following 12 classes were assumed: Nine nuclear tunes as used in the first part of the materials, two prenuclear tunes, H and L, and a class P, which is one or more unaccented syllables preceding a prenuclear tune.

For this part of the materials, 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 falling-rising.

1. $x_1 = V_p - F_e$

$x_1$ describes the difference between the initial value of $F_0$ (Vp) on the first vowel of the tune and the value of $F_0$ at the extremum (Fe).

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 ($F_{ming}$).

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. $x_7 = abs(V_p - F_k) - abs(V_{pa} - F_{ka})$

$x_7$ 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. $x_8 = (V_{pa} - F_{ka}) - abs(V_{ka} - F_{kr})$

$x_8$ is the difference between the variation of Fo 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.

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.

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.

In this second part of the materials, 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 with 6 hidden neurons.

Similar results were obtained for the classification of the 12 tunes with classical tree-layer MLP, the probabilistic and RBF. Detailed training of the MLP with three layers was performed. As usual, the number of input neurons corresponds to the number of features, which is 11, and the number of output neurons corresponds to the number of distinct classes, which is 12. The number of hidden neurons was varied between 10 and 20. Several cycles of training were performed with various h1 parameters (learning rate) and h2 parameters (momentum term).

1600 cases were used for training and 450 for testing. The cases were presented in random order. With 10 neurons the global error in training was 0.1021 and 0.1146 in testing. With 17 neurons the error in training was 0.067, and in testing 0.088 (Table 1). A further increase of the number of neurons did not improve the results. The scores were 82 percent correct both in training and in testing.

The poorest scores were obtained for the class P, the initial unaccented syllables: 67 percent in training and 60 percent in testing, and the best for LH, HL and LHL.

## 5. THE READ TEXTS

The materials in this experiment were fragments from 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.

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 and LxL formed the class of falling tunes, F.

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. The best results were obtained with 9 hidden neurons. The classification was, depending on the type of tune, between 79 and 83 percent correct.

The only, or the first syllable of every nuclear and prenuclear tune except P has pitch accent.

## 6. Conclusion

Polish accent is realized by intonation. There are, as probably in every intonation language, a finite number of melodic patterns, each pattern forming an intonation phrase.

This 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. A prenuclear tune may be preceeded by one or more syllables that do not form a strong prenuclear tune, and are therefore unaccented. Such unaccented syllables may also precede a nuclear tune directly. Nuclear and prenuclear tunes can be described in terms of the course of $F_0$ and the alignment of $F_0$ variation with syllabic cores.

By developing a relatively small number of mathematical formulations related to the $F_0$ course and its alignment with the syllables, it is possible to train artificial neural networks to detect and classify the various tunes in the spoken signal and by the same token to detect the pitch accent.

There are features of the speech signal that tend to signal the end of an intonation phrase (Sagisaka 1997), but an independent criterion categorically indicating the borders between intonations

groups has not been found. This may be the main reason for the occasional failures of the neural networks to detect and classify a tune correctly. A relatively high percentage of correct classification (i.e., in keeping with perceptual tests) confirms the utility of the parametrization method adopted here for the description of the intonation contour, and creates a basis for an extension of the training set 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 Fo 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 Tuned, 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.


| NNSET | KL | T_KL | | ERROR |
|---|---|---|---|---|
| 278 | L | L | Right | 1.191e-0 |
| 279 | L | L | Right | 1.747e-1 |
| 280 | HL | HL | Right | 0.000241 |
| 281 | L | P | Wrong | 0.4083 |
| 282 | L | L | Right | 5.042e-0 |
| 283 | L | L | Right | 1.323c-0 |
| 284 | HL | HL | Right | 1.037e-0 |
| 285 | H | H | Right | 4.321e-0 |
| 286 | ? | H | ? | 0.1355 |
| 287 | HL | HL | Right | 0.003271 |
| 288 | H | H | Right | 5.511e-0 |
| 289 | ML | ML | Right | 1.765e-0 |
| 290 | LH | LH | Right | 7.26e-07 |
| 291 | LH | LH | Right | 7.397e-1 |
| 292 | P | P | Right | 0.000246 |
| 293 | H | H | Right | 9.595e-0 |
| 294 | ? | H | ? | 0.257 |
| 295 | ? | H | ? | 0.2887 |
| 296 | ML | ML | Right | 5.16e-09 |
| 297 | ? | P | ? | 0.29 |
| 298 | L | L | Right | 0.01448 |
| 299 | ? | L | ? | 0.147 |
| 300 | L | L | Right | 0.01594 |
| 301 | HL | HL | Right | 0.001607 |
| 302 | ML | ML | Right | 7.224e-0 |
| 303 | ML | ML | Right | 3.02e-05 |
| 304 | ML | ML | Right | 5.003e-1 |
| 305 | LM | LM | Right | 3.194e-0 |
| 306 | LM | LM | Right | 4.929e-0 |
| 307 | LM | LM | Right | 5.787e-0 |
| 308 | ML | P | Wrong | 0.4082 |
| 309 | ML | ML | Right | 0.002355 |
| 310 | ML | Ml | Right | 0.00273 |
| 311 | H | H | Right | 0.000445 |
| 312 | L | L | Right | 0.00051 |

Table 1. Extract from a classification output.
KL-expected class, T_KL- observed class E_KL-result
Error- mean-square error of classification

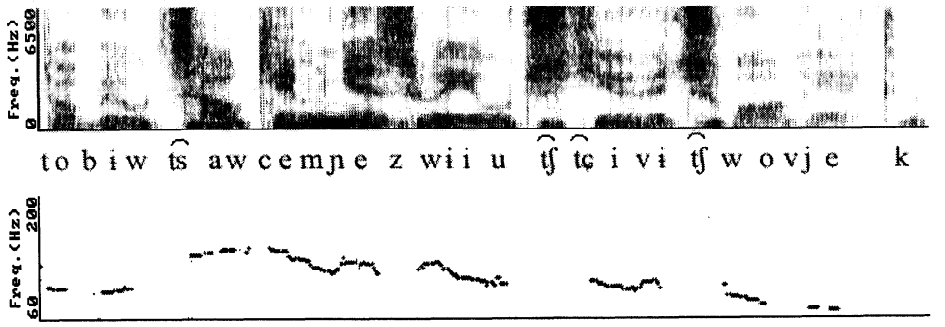


Fig.1a.

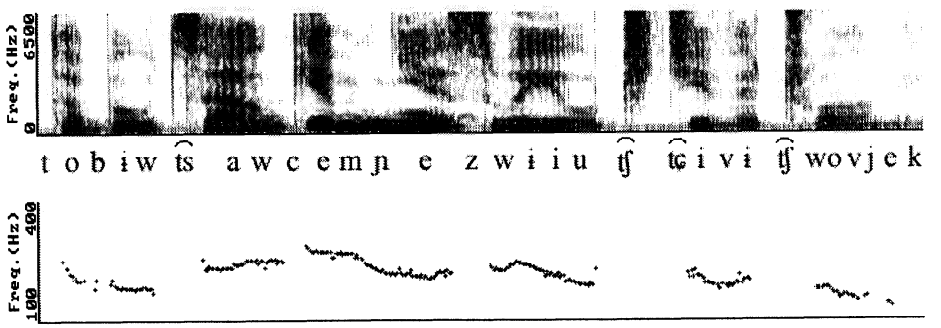


Fig.1b.


Fig.2a.

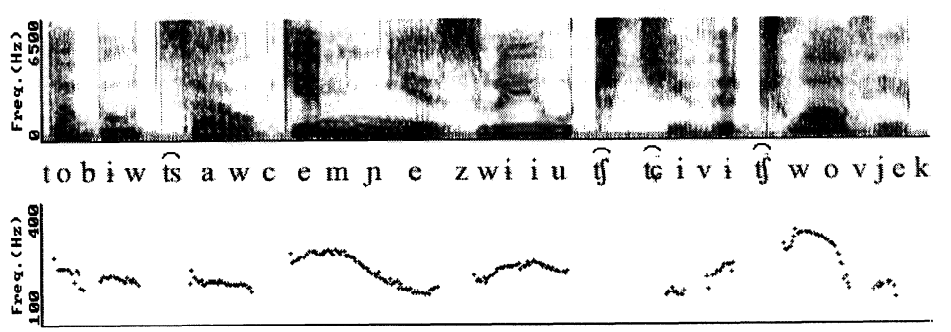


Fig.2b.

Fig. 1a and 1b, 2a and 2b.
The phrase is (1) *To był* (H)*całkiem* (H)*niezły i ucz*(H)*ciwy* (ML)*człowiek.* (He was quite a good and honest man) with three H prenuclear tunes and and ML nuclear tune, and (lexically) the same phrase with three L prenuclear tunes: (2) *To był* (L)*całkiem* (L)*niezły i ucz*(L)*ciwy* (ML)*człowiek.*