ARCHIVES OF ACOUSTICS
23, 2, 159–178 (1998)

ACOUSTIC CLASSIFICATION AND AUTOMATIC BORDER OF PHRASE AND STRESS RECOGNITION IN POLISH

G. DEMENKO

Institute of Linguistics
A. Mickiewicz University
(60-371 Poznań, Międzychodzka 5, Poland)

A classification of perceptively and linguistically verified intonation structures was made basing on acoustic qualities of speech signals. The importance of the eight selected parameters describing the set of elements under investigation was tested by means of statistical methods. Findings of discriminant analysis proved the possibility of a correct classification of unstressed vowels (in the range of 86 – 92%). For the remaining vowels (stressed and located at the end of a phrase) a significant degree of correct classification was obtained in the range of 54 – 73%. Far better classification results (88% at the average) were obtained by using a neural network. Identification of accents and phrase boundaries originating from the new data set showed the necessity to extend the linguistic base and repeat the network training.

1. Introduction

The automatic extraction, analysis and synthesis of suprasegmental features of speech is indispensable in computer systems of verbal communication. One of the most difficult problems requiring practical solution is the division of a continuous speech signal into smaller units (cf. e.g. [5, 7, 15]). In a written text, the segmenting function is performed by punctuation marks whose presence enables the recipient to carry out the division of the text into information units according to the intention of the sender. In an oral text, the division of the text into units called phrases is achieved by the realisation of specific intonation patterns, the rhythmic structure and the presence of pauses.

In the work [3] an experiment to determine the possibility of establishing perceptive phrase boundaries and the place of the stress occurrence was carried out. The experiment testified a high degree of uniformity among listeners as regards the segmentation of a speech signal. The listeners marked altogether (in three versions of the text) 221 phrase boundaries and 352 stressed syllables in the material under analysis (a three minute press article read out by three speakers: two male voices and one female voice). A linguistic analysis of the text assessed perceptively was carried out (cf. [10]).

The syntactic cohesion of the phrase was tested by embedding unilaterally (to the left) successive accentual-syntactic units. A perceptively distinguished phrase displays

grammatical coherence in the form of a specific syntactic structures and phonetic coherence in the form of a specific stress distribution.

The findings of the perceptive-linguistic analysis were used for preparing the acoustic classification of intonation structures of Polish in read out texts. In the paper presented now priority was granted to the formulation and statistical testing of the qualities describing suprasegmental structures of speech at the acoustic level. Two acoustic parameters of the signal were subjected to a detailed analysis, i.e. the vowel duration and the fundamental frequency alterations. Relative alterations of an F_0 parameter in the adjacent syllables play an important part (cf. [9, 12]) and the vowel duration is of significant importance in the division of the speech signal into phrases (cf. e.g. [11]). Variations in the intensity of stress perception are of less significance than alterations of the fundamental frequency and vowel duration. Therefore the analysis of this parameter was disregarded. A measurement of the signal level was mostly used when segmenting the speech signal into syllables.

A digital Kay spectrograph and specialised CSRE 4.5 software, which makes the extraction of significant parameters of the speech signal possible, were used in this work.

2. Acoustic analysis of suprasegmental structures

2.1. Measurement of the vowel duration

Publications on vowel duration in Polish are very scarce and comprise only most rudimental issues (cf. [6]). It was assumed that acoustic parameters within adjacent vowels play a major role in the perception of suprasegmental qualities. There is no automatic method of determining vowel boundaries fairly correctly (i.e. with an error comparable to the inaccuracy of the manual measurement). Obtaining sufficiently accurate measurements of the vowel duration time is a necessary condition of the duration analysis. Therefore, a phonetician-operator determined manually the vowel boundary basing on the observation of the time pattern, the signal spectrum and simultaneous sound reproduction. For some sound clusters, the determination of the boundary was virtually impossible. This was the case with vowel clusters with /j/ or /w/, /l/ or nasal consonants. Such inseparable fragments of the signal were treated as diphthongs. A specific case was constituted by inseparable clusters of three sounds (e.g. /eja/, /ow/, etc.) which were classified as triphthongs. The total number of vocalic segments extracted from the text was 599, 595 and 537 in the respective voices. The classification of the segments eventually assumed for the whole experimental material (three voices) is based on accounting four possibilities, i.e.:

Classification I: phonetic segment duration proper
classes: 1 – short vowel, (i ɨ u), 2 – long vowel (e a o), 3 – diphthong, 4 – triphthong.

Classification II: stress placement
classes: 1 – unstressed vowel, 2 – stressed vowel.

Classification III: position of a syllable in a phrase

classes: 1 – syllable other than ultimate or penultimate in a phrase, 2 – stressed penultimate syllable in the phrase, 3 – stressed ultimate syllable in the phrase.

These three classes pertained to monophthong segments. In the case of diphthongs and triphthongs, two classes were distinguished: class 2 pertained to both the ultimate and penultimate syllable in the phrase.

Classification IV: structure of the ultimate syllable in the phrase

classes: 1 – closed syllable (ending with a consonant), 2 – open syllable (ending with a vowel).

The manner of classification assumed made it possible to distinguish 20 classes of vocalic segments occurring in the material under examination (Table 1).

Table 1. Classes of vocalic segments.

Class No	Vowel	Syllable
1	short unstressed	not ultimate or penultimate in the phrase
2	long unstressed	not ultimate or penultimate in the phrase
3	short stressed	not ultimate or penultimate in the phrase
4	long stressed	not ultimate or penultimate in the phrase
5	long stressed	not ultimate or penultimate in the phrase
6	long unstressed	ultimate, closed
7	short stressed	ultimate, open
8	short unstressed	ultimate, closed
9	long unstressed	ultimate, open
10	short unstressed	ultimate, closed
11	short stressed	ultimate, closed
12	long stressed	penultimate
13	stressed diphthong	penultimate
14	stressed diphthong	penultimate
15	stressed unstressed	not ultimate or penultimate
16	stressed unstressed	ultimate
17	stressed triphthong	not ultimate or penultimate
18	stressed triphthong	penultimate
19	stressed unstressed	not ultimate or penultimate
20	stressed unstressed	ultimate not ultimate or penultimate

It results from the four classifications accounted for, that the greatest influence on the duration of the vowels is exerted by their position in a phrase. Prolongation of

vowels in two last syllables should be considered a characteristic feature which signals the occurrence of a phrase boundary.

Figure 1 presents the value of average duration in classes normalised for all the voices (x_i means average duration in class i). The data from the chart render the observation of the influence of individual variables upon vowel duration possible. Hence the mean values for short vowels are lower than those for the long ones at the same conditions: ($x_1 < x_2$, $x_3 < x_4$, $x_7 < x_5$, $x_{11} < x_{12}$); the mean values for unstressed vowels are lower than those for the stressed ones: ($x_1 < x_3$, $x_2 < x_4$, $x_9 < x_5$, $x_{10} < x_7$, $x_{16} < x_{14}$, $x_{20} < x_{18}$); the mean values for vowels in the penultimate or ultimate syllable are higher than those for vowels in the remaining syllables: ($x_1 < x_8$ and x_{10}, $x_2 < x_6$ and x_9, $x_3 < x_7$ and x_{11}, $x_4 < x_5$ and x_{12}, $x_{14} < x_{13}$, $x_{16} < x_{15}$, $x_{18} < x_{17}$, $x_{20} < x_{19}$) and the mean values for vowels in the closed ultimate syllable are lower than those for vowels in the open syllable: ($x_9 < x_6$, $x_{10} < x_8$).

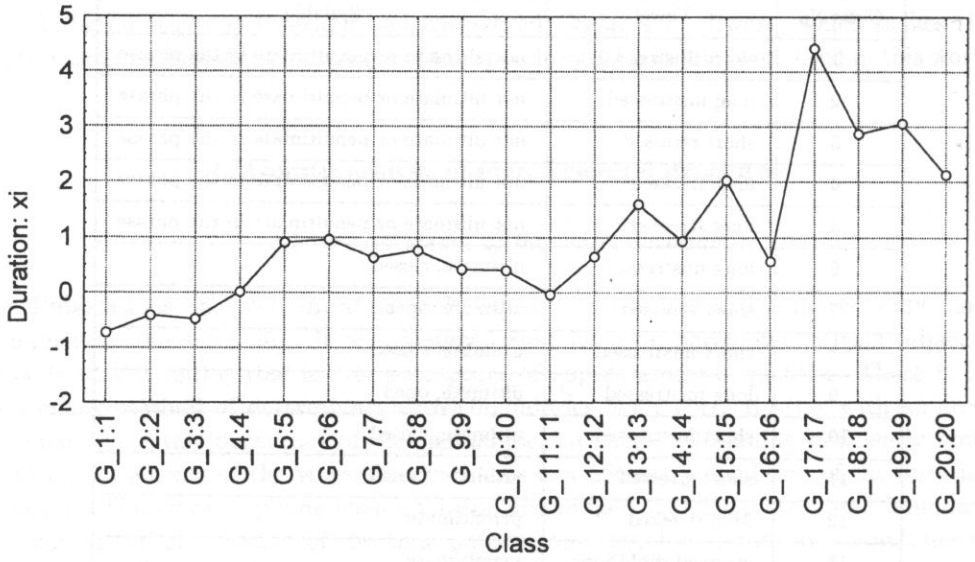

Fig. 1. Normalized means of vowel duration $F(19, 1711) = 131.06$, $p < 0.000$. Voices: LR, JI, MG.

The normalisation of the vowel duration, according to theoretical premises, makes it necessary to allow for a number of factors shaping the time structure of an utterance (cf. [1, 8, 11]) and is practically very difficult to carry out. Since in the texts under examination the greatest influence on vowel duration was exerted by the prolongation of the final sounds of the phrase, it can be assumed that the disregarding other factors is of little importance with respect to the classification of the vowel duration dependent on the syllable position. For each voice examined separately, vowel durations were normalized to the value whose distribution is characterised by a zero average value and an elementary standard deviation (z transformation). Attempts of normalization of this type were carried out (mostly for practical applications) also in other languages (cf. e.g. [1]).

2.2. Variability of the fundamental frequency

Intonation structures of read-out texts verified perceptively were subjected to an acoustic analysis. Using a packet of speech analysis programmes, the signal segmentation was carried out which made it possible to assign the frequency alterations of the fundamental frequency recorded instrummentally to selected phonetic units – vowels (cf. [4]).

All syllables occurring in the text were subject to an acoustic analysis. Several syllables for which it was practically impossible to define acoustic parameters (most often they were of low amplitude in the final position syllables), were excluded from the statistical analysis. In the majority of cases, the F_0 parameter pattern on stressed vowels/syllables was of rising or constant character. Patterns in which the fluctuations of the F_0 parameter did not exceed 15 Hz were considered to be constant. Falling or rising and falling patterns occurred more rarely. Extreme values of the F_0 parameter on stressed syllables usually occurred near the global extremes of the pattern (most often maxima). In several instances a minimum value of the parameter on a stressed vowel (not located in the vicinity of the phrase boundary), constituting a local minimum of the pattern, was observed. The identification of such a stress can be possible by the analysis of the duration of individual vowels and the local minimum of the F_0 pattern. Stressed vowels in a structure of this type were observed to be not shorter than the adjacent unstressed ones.

For all syllables and vowels occurring in the text, the variability of the fundamental frequency was described by a separate set of parameters describing the initial, final, maximum and minimum values of subsequent vowels ($V_{\min}, V_{\max}, V_p, V_k$) and consonants ($F_{\min}, F_{\max}, F_p, F_k$) with other derived parameters constituting their combinations. A created set of 28 selected variables for vowels is presented in Annex 1.

3. Statistical analysis

Basing on the experimental material, four classification options were taken into account (similar division was assumed for syllables):

CL1 – classification 1
- class 1 – (cl1) – unstressed vowels
- class 2 – (cl2) – stressed vowels
- class 3 – (cl3) – ultimate vowels in a phrase – with rising intonation
- class 4 – (cl4) – ultimate vowels in a phrase – with falling intonation
- class 5 – (cl5) – vowels preceding vowels from class 3
- class 6 – (cl6) – vowels preceding vowels from class 4

CL2 – classification 2
- class 1 – (cl1) – unstressed vowels
- class 2 – (cl2) – stressed
- class 3 – (cl3) – ultimate before the phrase boundary

CL3 – classification 3
– class 1 – (cl1) – unstressed vowels
– class 2 – (cl2) – stressed or two ultimate before the phrase boundary
CL4 – classification 4
– class 1 – (cl1) – vowels not placed before the phrase boundary
– class 2 – (cl2) two ultimate vowels before the phrase boundary.

The data were normalized by deducting the logarithm of a minimum value (characteristic for a given voice and determined as an average value of the F_0 parameter values occurring at the end of affirmative sentences) from the logarithms of subsequent values of the F_0 parameter. Abiding the requirements of further data analysis, such normalized values were located in a changing range from -1 to 1.

The analysis of variance showed the importance of the 28 selected qualities for specific classifications. It turned out that the statistical characteristics describing the variability of the F_0 parameter within a syllable or vowel (e.g. minimum, maximum values, scope of alterations) are of little use for statistical classification. More important are dynamic parameters, i.e. those which describe the relations in the adjacent syllables (e.g. the ratio of the extreme values of F_0 in a current and preceding syllable). Parameters describing vowels (not syllables) proved to be more important.

Therefore, algorithms for the recognition of stressed syllables based on statistical averaging of extreme values of vowels (or syllables) and, consequently, of classifying intonation patterns are not effective.

In the selected set of qualities a strong correlation between some qualities were observed, e.g. between the initial and final values of the F_0 parameter of a vowel are strongly correlated within the vowel with maximum and minimum values (the patterns of vowels usually do not have an extreme). Annex 2 presents an example of findings from the analysis of variance for vowels described by the set of 28 qualities. Combinations of features (not correlated to each other) which allow for the duration of a vowel, differential between extreme values of the F_0 parameter and the sum of absolute changes of F_0 upon adjacent sounds, were selected for vowel classification.

Figure 2 illustrates average values, Fig. 3 illustrates the medians and the range of variations of the specific qualities (Figs. 3.1–3.8) according to the **CL1** classification. The following interpretation was assumed for the eight variables selected:

1. DUR: – duration

This quality groups the vowel durations (Fig. 3.1) in the following order: cl1, cl2, cl6, cl5, cl3, cl4. The biggest differences of the statistical parameters were observed for vowels from class cl1 – unstressed (over 75% of the data exceed the value of -0.6) and those from the classes cl3 and cl4 situated immediately before the phrase boundary (over 75% of the values exceed -0.5). Vowel durations from the remaining classes assume medium values.

2. DMIN $= V_{\min} - V_{\min -1}$

The DMIN variable depitcs the difference between the minimum value of the F_0 parameter on a current vowel and a minimum value thereof on the preceding vowel.

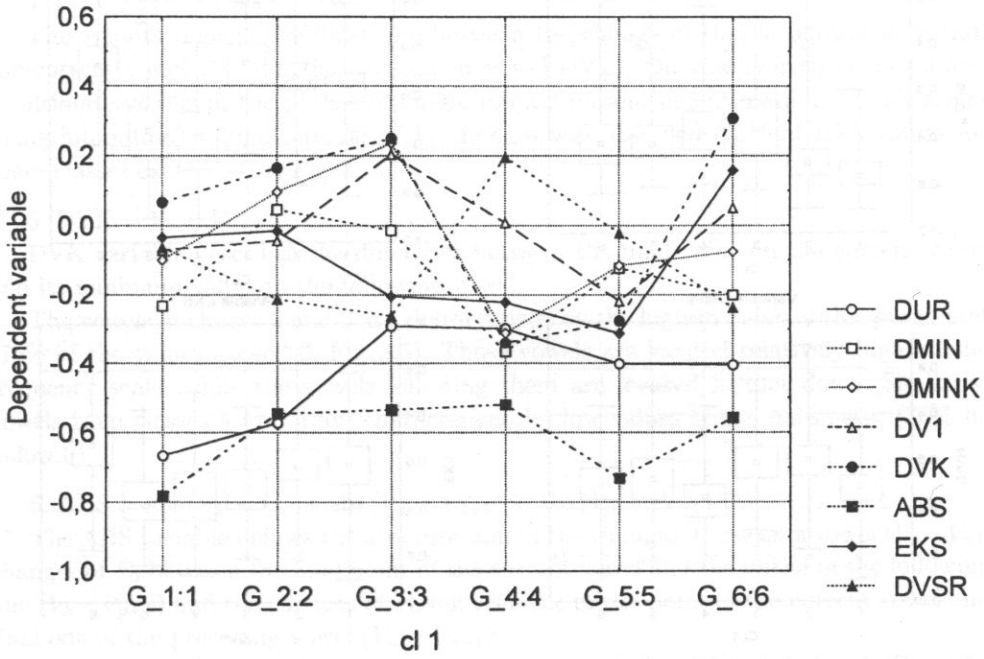

Fig. 2. Means of 8 features according to classification 1.

The lowest values of the parameter (Fig. 3.2) were observed in class 4. For vowels of classes 2 and 3 the DMIN parameter assumes the highest values (75% of the data exceed -0.2). The result is justified: a vowel at the end of an utterance of falling (cl4)/rising (cl5) intonation is located not higher/lower than the preceding vowel. In the case of stressed vowels (cl2), a minimum value of the F_0 parameter on a current vowel may be higher or lower than the minimum value on the preceding vowel. In a read-out text only in very few cases a minimum value was observed on the stressed vowel which constitutes a local minimum of the pattern. Therefore the DMIN parameter in the set of data examined achieves high values for stressed vowels.

3. DMINK $= 2 * V_k - V_{\min -1}$

The DMINK parameter depicts the sum of the final F_0 parameter value on a current vowel (V_k) and the value of the difference between the final value on the current vowel and minimum value on the preceding vowel ($V_k - V_{\min -1}$). This parameter distinguishes vowels of classes 4 and 3 (Fig. 3.3). V_k is always low in class 4 – it constitutes a global minimum of the pattern in the phrase (in the case of laryngalisation it often falls below the statistical $F_{\min}$). The final value of F_0 is always high in class 3 (it constitutes a global extreme of the pattern or is located in its vicinity). If both components of the sum: $V_k - V_{\min -1}$ and V_k are small, their sum is also small and indicates that the vowel belongs to class 4. In the case of vowels from classes: 2, 5, 6 the value of DMINK parameter, due to higher V_k, will not achieve such low values as in class 4.

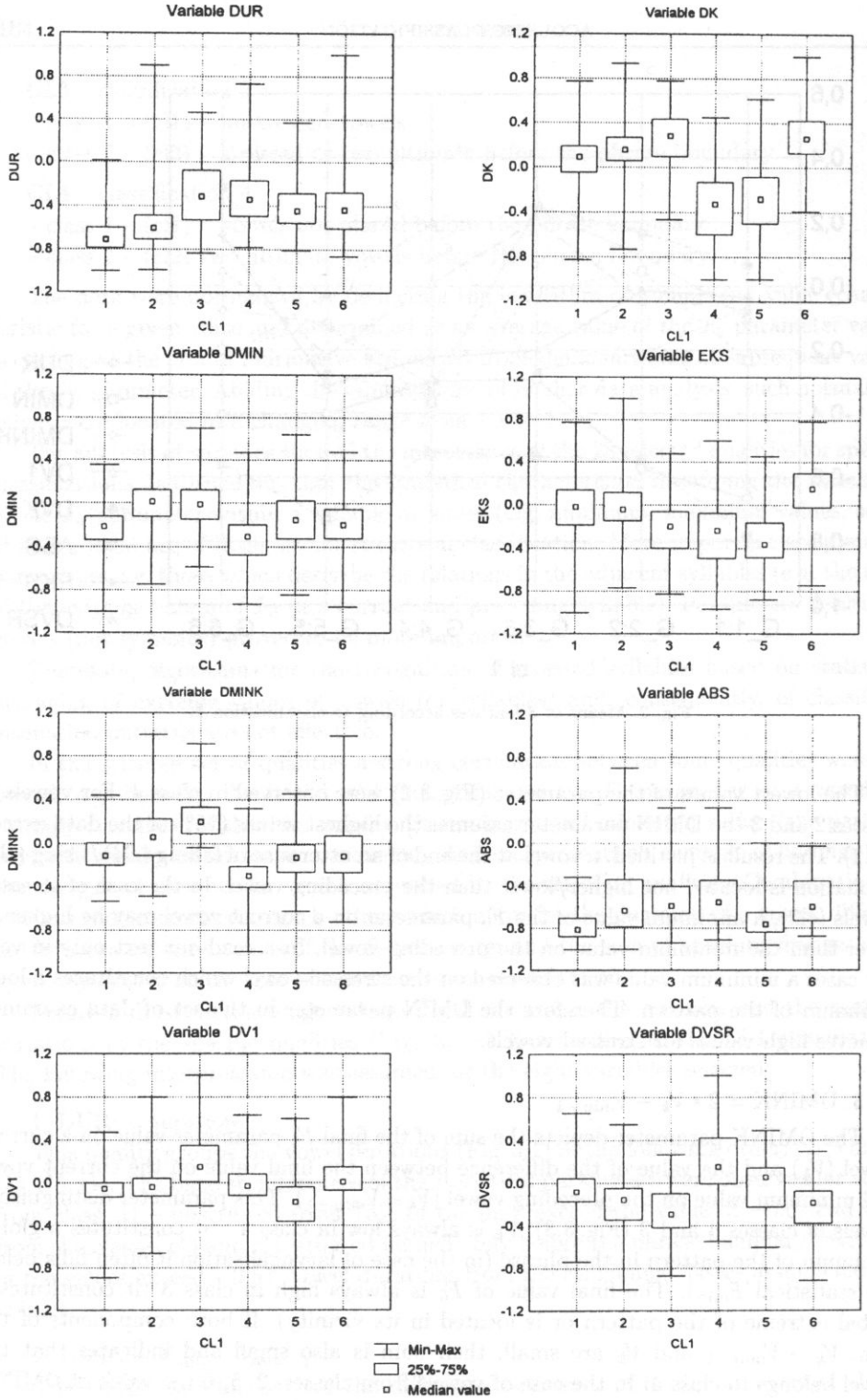

Fig. 3. Range of variables for classification 1.

4. $\mathrm{DV1} = \mathrm{DV} - \mathrm{DV}_{+1}$

This quality depicts the difference between the change of the F_0 parameter within the current vowel DV and the subsequent vowel DV_{+1}. On vowels from classes 3 and 6 (ultimate vowels in the phrase of falling intonation and penultimate in the phrase of rising intonation) a high interval of F_0 changes was observed (75% of DV1 values for these classes exceed -0.1, Fig. 3.4).

5. $\mathrm{DVK} = V_k - V_{k+1}$

DVK variable describes the difference between the final value on the current vowel and its minimum value on the following vowel.

The vowels in classes 6 and 3 are characterized by the highest value of this parameter (75% of the values exceed 0, Fig. 3.5). Those vowels are located relatively high on the frequency scale, while the vowels following them are located further down. Similarly, vowels from classes 4 and 5 are characterized by low values of the parameter (75% lie below 0).

6. $\mathrm{ABS} = \mathrm{abs}\,(V_k - V_p) + \mathbf{abs}\,(V_k - V_{p+1}) + \mathrm{abs}\,(V_p - V_{k-1})$

The ABS variable defines the absolute sum of the changes: those on a vowel $(V_k - V_p)$, changes of F_0 between the final point of the current vowel and the initial of the following one $(V_k - V_{p+1})$ and the changes of F_0 between the initial point of the current vowel and final one of the preceding vowel $(V_p - V_{k-1})$.

The biggest changes of the parameter occur on vowels from classes 2, 3, 4, 6 (Fig. 3.6). On unstressed vowels either small changes in the value of the F_0 parameter occur or the initial/final values are close to the values on the adjacent vowels.

7. $\mathrm{EKS} = (V_{k+1} - V_{p+1}) - (V_k - V_p) - (V_{p+1} - V_k)$

The EKS parameter defines the difference of frequency alteration in two adjacent vowels $(V_{k+1} - V_{p+1})$ and $(V_k - V_p)$ and the change of values between the initial point of the following vowel and the final point of the current vowel $(V_{p+1} - V_k)$.

For vowels of class 6, the EKS parameter assumes the highest values (75% of the data exceed 0), for vowels from class 5 the parameter assumes low values (75% lie below -0.2).

8. $\mathrm{VSR} = (\mathrm{VSR}_{-1} - \mathrm{VSR}) - (\mathrm{VSR} - \mathrm{VSR}_{+1})$

An analogy may be observed between the VSR variable defining the differential of differences between the mean values on the current and preceding vowel ($\mathrm{VSR}_{-1} - \mathrm{VSR}_{+1}$) as well as between the current and the following one ($\mathrm{VSR} - \mathrm{VSR}_{+1}$) to the second derivative at the point specified by VSR (average value on the current vowel).

For vowels with maximum value of F_0 (class 3 and 6), the pattern approximating the change between adjacent vowels is convex ($\mathrm{VSR} < 0$), for patterns with a minimum (vowels from class 4) they are concave ($\mathrm{VSR} > 0$). This variable differentiates the vowels of classes 3 and 6 (75% of the data lie below -0.2) from those of class 4 (75% of the data below 0).

A discriminant analysis carried out on a selected set of patterns of the fundamental frequency of Polish ([2]) proved the possibility for a statistical classification of the intonation structures. The discriminant analysis was carried out for four selected classifications: **CL1**, **CL2**, **CL3**, **CL4**.

Table 2a presents the results of the analysis for the classification **CL1**. Unstressed vowels were classified relatively well (in 90% of the cases). In other groups the percentage of correct classifications oscillates around 50%–60%. For class 6 only 32% of the classifications was correct. Figure 4 illustrates vowels ordered in the co-ordinating system of the first two discriminant variables. It can be observed that the classes are linearly inseparable and posses a complex geometrical structure. Unstressed vowels (class 1) can be separated from the others relatively good (Fig. 4.1).

Table 2a. Results of the discriminant analysis for the **CL1** classification.

		Classifications					
Class	Percentage of compatibility	cl1	cl2	cl3	cl4	cl5	cl6
cl1	90	759	36	0	4	12	7
cl2	56	156	247	21	2	7	8
cl3	57	16	25	67	2	0	7
cl4	51	27	2	0	46	13	1
cl5	49	37	14	2	4	56	0
cl6	32	39	23	9	4	0	36

Table 2b. Results of the discriminant analysis for the **CL2** classification.

		Classifications		
Class	Percentage of compatibility	cl1	cl2	cl3
cl1	90	741	35	41
cl2	55	154	245	43
cl3	54	125	72	233

Table 2c. Results of the discriminant analysis for the **CL3** classification.

		Classifications	
Class	Percentage of compatibility	cl1	cl2
cl1	85	706	111
cl2	73	230	639

Table 2d. Results of the discriminant analysis for the **CL4** classification.

		Classifications	
Class	Percentage of compatibility	cl1	cl2
cl1	93	1186	72
cl2	46	230	198

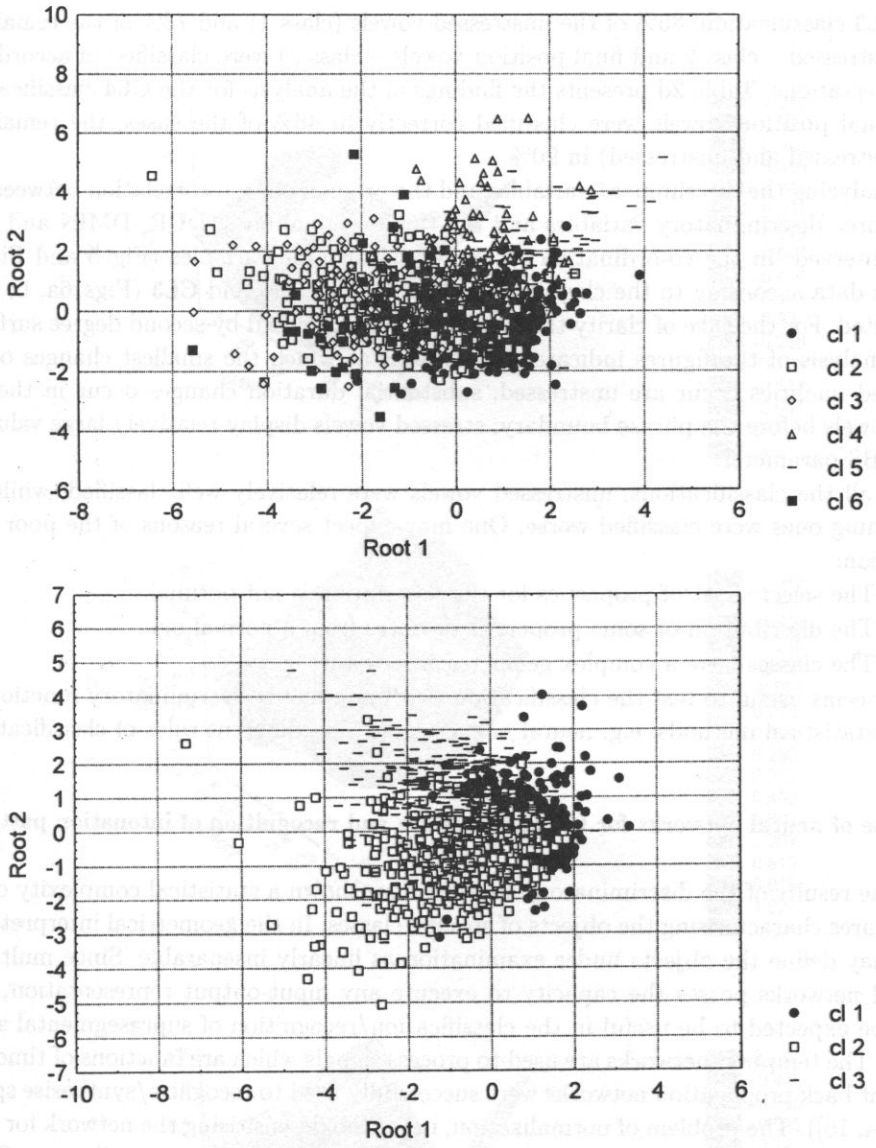

Fig. 4. The vowels in a discriminant-variables plane.

Table 2b presents the results of the analysis for classification **CL2**. A correct classification was obtained for 55% vowels from class 2 and 54% of those of class 3. Unstressed vowels were classified correctly in 90% of the cases. Figure 4.2 presents the data in the co-ordinating system of the first two discriminatory variables for the **CL2** classification.

The first discriminatory variable separates unstressed vowels (class 1) from the remaining ones, the second variable differentiates stressed vowels and vowels occurring in the final position of the phrase (class 2). Table 2c presents the results of the analysis for

the **CL3** classification. 85% of the unstressed vowels (class 1) and 73% of the remaining ones (stressed – class 2 and final position vowels – class 3) were classified in accordance to expectations. Table 2d presents the findings of the analysis for the **CL4** classification. The final position vowels were classified correctly in 46% of the cases, the remaining ones (stressed and unstressed) in 90%.

Analyzing the discriminant variables and the original ones, a correlation between the first three discriminatory variables and the three original ones: DUR, DMIN and ABS was observed. In the co-ordination system of those three variables (Fig. 5 and Fig. 6), all the data according to the classification **CL2** (Figs. 5 a–c) and **CL3** (Figs. 6a, b) were presented. For the sake of clarity the data were approximated by second degree surfaces. The analysis of the figures indicates that vowels at which the smallest changes of the selected qualities occur are unstressed; substantial duration changes occur in the last two vowels before the phrase boundary; stressed vowels display relatively large values of the ABS parameter.

In all the classifications, unstressed vowels were relatively well classified, while the remaining ones were classified worse. One may expect several reasons of the poor classification:

1. The selected set of properties for the description is not optimal.
2. The distribution of some properties deviates from a normal one.
3. The classes have a complex geometrical structure.

It seems useful to test the classifications with non-linear discriminatory functions or other statistical methods, e.g. neural networks utilising different rules of classification.

4. Use of neural networks for the classification and recognition of intonation patterns

The results of the discriminatory analysis have shown a statistical complexity of the structures characterising the objects of specific classes. In the geometrical interpretation one may define the objects under examination as linearly inseparable. Since multilevel neural networks posses the capacity to execute any input-output representation, they may be expected to be useful in the classification/recognition of suprasegmental structures. The temporal networks are used to process signals which are functions of time. Recurrent back propagation networks were successfully used to recognise/synthesise speech (cf. [14, 16]). The problem of normalization, i.e. of the desensitising the network for some time transformations of the input images (e.g. change of the speech speed) is not finally solved in temporal networks. Since selected qualities of suprasegmental structures reflect time relations in three adjacent syllables (in the majority of cases, disregarding complex structures, of *gradual rise* or *gradual fall* type), such a description suffices for the purposes of automatic classification. For the classification of examined structures a classical three-layer hetero-association back propagation network utilising the momentum element (Fig. 7) was used. Subsequent layers were joined "with each other". Hidden and input layers were additionally joined with "constant 1" type elements (bias). The number of input neurals is defined by the dimensions of the input vector (eight qualities describing specific intonation structures). The **Cl2** classification (stressed vowel, unstressed, ulti-

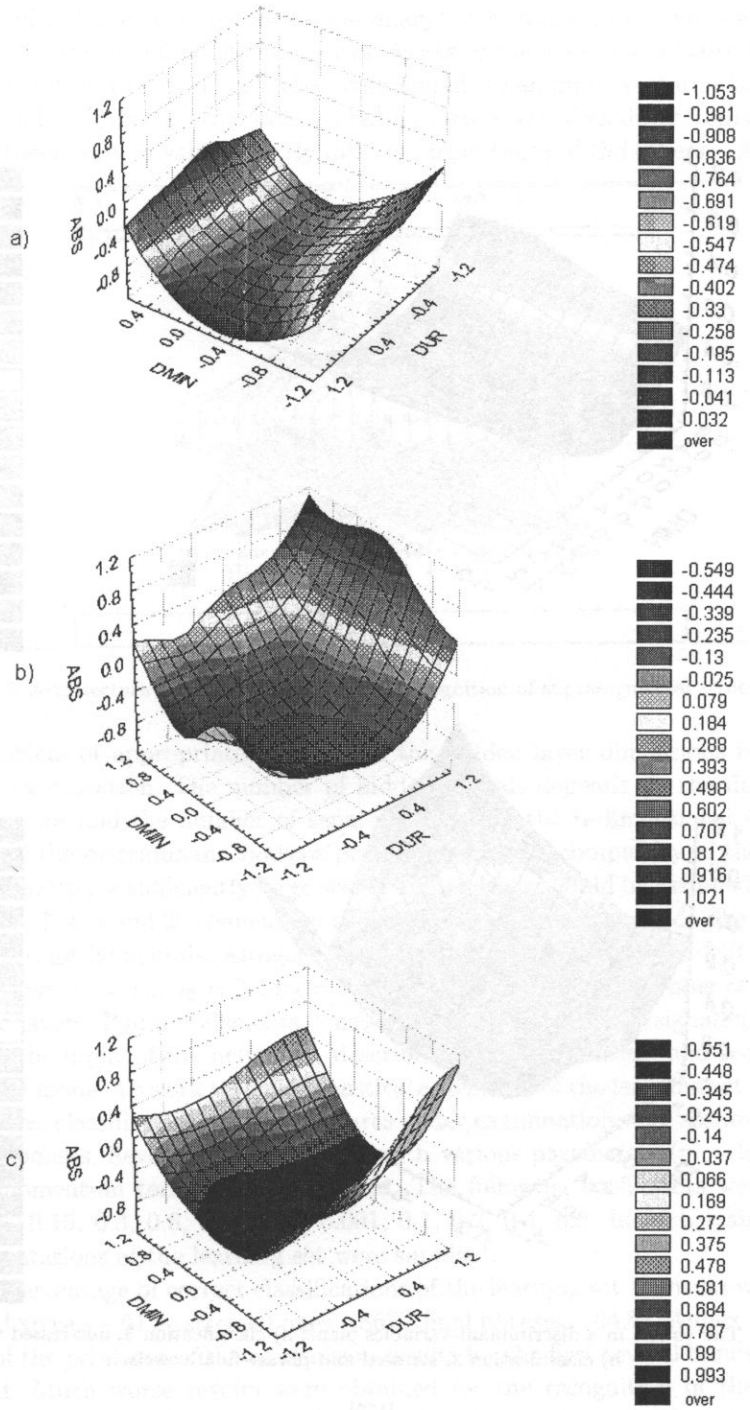

Fig. 5. The vowels in a discriminant-variables plane; a) classification 2, unstressed vowels, b) classification 2, stressed vowels, c) classification 2, phrase-final vowels.

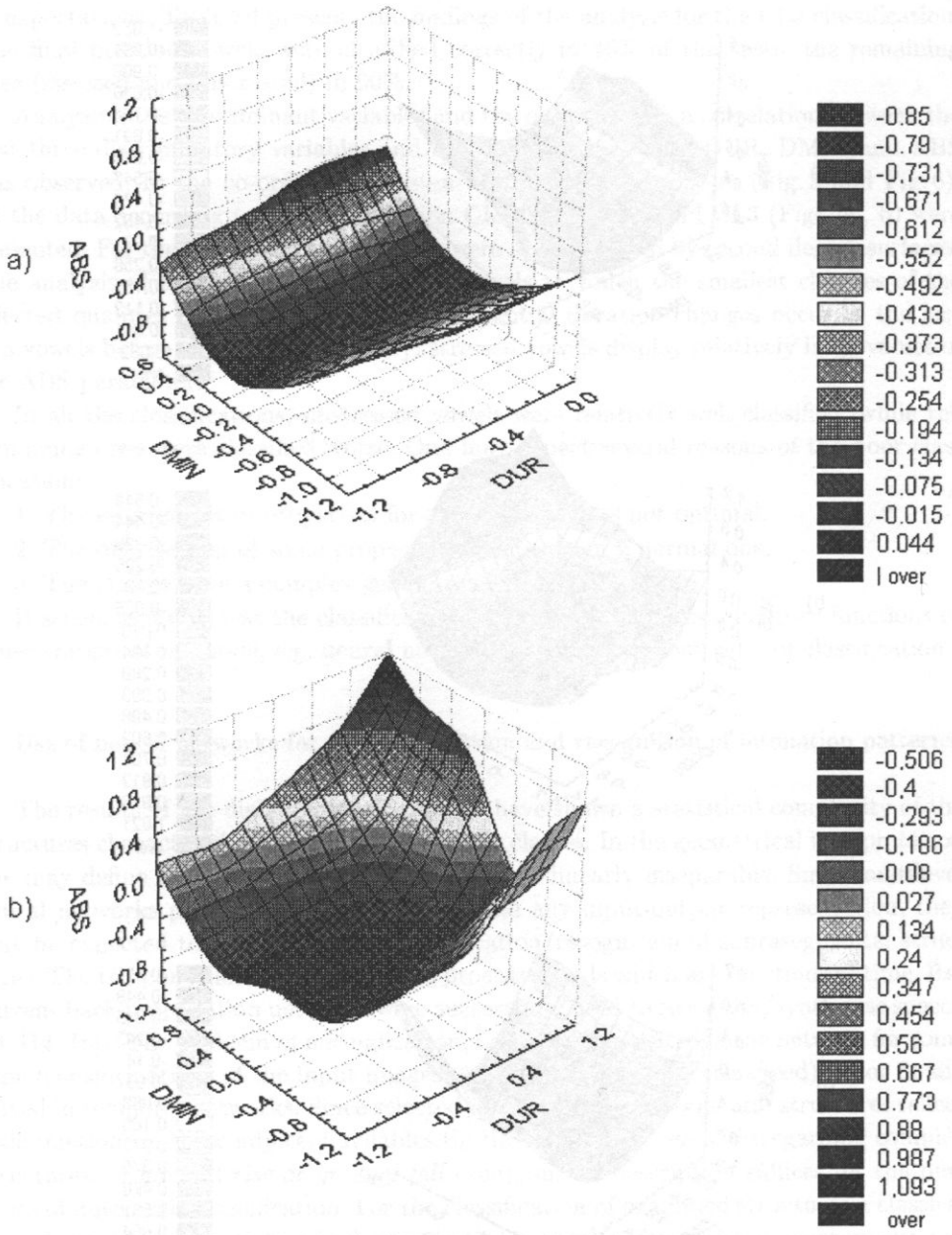

Fig. 6. The vowels in a discriminant-variables plane; a) classification 3, unstressed vowels, b) classification 3, stressed and phrase-final vowels.

mate or penultimate before the phrase boundary) was assumed. The number of neurals in the input layer is equal to the number of classes in the case of a network classifier. A local representation was used; each class corresponds to an input vector in that only one component differs from the other ones (in reality all of them often differ from one another but one of them accepts values clearly different from those of the others – Annex 3).

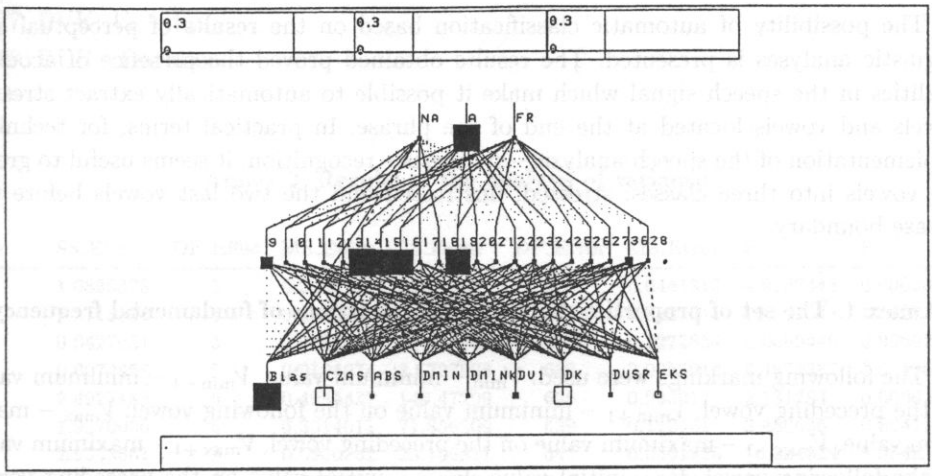

Fig. 7. Architecture of the neural net for the recognition of suprasegmental structures.

The problem of appropriate selection of the hidden layer dimensions is subject to continuous examination. The number of hidden neurals depends on the dimensions of the input vector and the number of separable areas in the n-dimensional space. Since the results of the discriminant analysis proved geometrical complexity of the structures under examination, a sufficiently large size of hidden layer should be assumed. A network learning with 2, 4, 8 and 20 elements in the hidden layer was conducted. The best results were obtained for 20 neurals. Altogether $(n_1 + 1) * n_2 + (n_2 + 1) * n_3 = 240$ connections were left, where $n_1 = 18$, $n_2 = 20$, $n_3 = 3$ stand for the respective number of elements in the specific layers. Pairs of elements consisting of vectors of input signals (information supplied at the input of the network – described by eight qualities) and required input signals of the model network response constitute elements of the learning set. The results of a perceptive classification of the structures under examination were assumed as model network responses. Several learning cycles with various parameters (η_1 – learning rate and η_2 – momentum term) were conducted. The following coefficients were assumed: $\eta_1 = 0.015, 0.15, 0.3, 0.6, 0.9$; $\eta_2 = 0.001, 0.1, 0.2, 0.4, 0.6$. In each training session 20000 presentations of the learning set were supplied.

A high percentage of correct classifications of the learning set elements was obtained (unstressed vowels – 91%, stressed ones – 86%, final phrases – 84%). Annex 3 contains a fragment of the print-out of the recognition results for the first several hundred syllables of the text. Much worse results were obtained for the recognition of the intonation structures than those of their classification. Vowels not from the learning set (from different language material) were correctly recognised in 75% at the average. This result

indicates the necessity of completing the learning (training) material, i.e. integrating the two data sets and training the network again.

5. Conclusions

The possibility of automatic classification based on the results of perceptual and linguistic analyses is presented. The results obtained proved the presence of acoustic qualities in the speech signal which make it possible to automatically extract stressed vowels and vowels located at the end of the phrase. In practical terms, for technical implementation of the speech analysis – synthesis – recognition, it seems useful to group the vowels into three classes: stressed, unstressed and the two last vowels before the phrase boundary.

Annex 1. The set of properties describing the variations of fundamental frequency

The following markings were used: $V_{\min}$ – minimum value, $V_{\min -1}$ – minimum value on the preceding vowel, $V_{\min +1}$ – minimum value on the following vowel; $V_{\max}$ – maximum value, $V_{\max -1}$ – maximum value on the preceding vowel, $V_{\max +1}$ – maximum value on the following vowel, V_p – initial value, V_{p-1} – initial value on the preceding vowel, V_{p+1} – initial value on the following vowel, V_k – final value, V_{k-1} – final value on the preceding vowel, V_{k+1} – final value on the following vowel, V_s – average value, V_{s-1} – average value on the preceding vowel, V_{s+1} – average value on the following vowel.

1. V1: $V_{\min}$
2. V2: $V_{\max}$
3. V3: $V_{\max} - V_{\min}$
4. V4: $DV - DV_{-1}$
5. V5: $V_{\min} - V_{\min -1}$
6. V6: $V_{\max} - V_{\max -1}$
7. V7: $V_{\max} - V_{\max +1}$
8. V8: V_{sr}
9. V9: $V_{sr} - V_{sr-1}$
10. V10: $V_{sr} - V_{sr+1}$
11. V11: V_p
12. V12: V_k
13. V13: $V_k - V_{k-1}$
14. V14: $V_k - V_{k-2}$
15. V15: V_{k-1}
16. V16: $V_{k-1} - V_{p-1}$
17. V17: $V_{k+1} - V_{p+1}$
18. V18: $V_{\max} - V_k$
19. V19: $V_{\max} - V_k$
20. V20: $V_{\max -1} - V_k - V_{\max -3} - V_{k-2}$
21. DMINK: $2 * V_k - V_{\min -1}$

22. ABS: $\text{abs}(V_p - V_{k-1}) + \text{abs}(V_k - V_{p+1}) + \text{abs}(V_k - V_p)$
23. EKS: $(V_{k+1} - V_{p+1}) - (V_{p+1} - V_k) - (V_k - V_p)$
24. VSR: $(V_{sr-1} - V_{sr}) - (V_{sr} - V_{sr+1})$
25. DV1: $D_v - D_{v+1}$
26. DMIN: $V_{\min} - V_{\min -1}$
27. DVK: $V_k - V_{k+1}$
28. DUR : Duration

Annex 2. Results of the analysis of variance

Case	SS_Effec	DF_Effec	MS_Effec	SS_Error	DF_Error	MS_Error	F	P
V1	1.0835873	5	0.2167175	28.244021	640	0.0441313	4.9107448	0.0002031
V2	1.2718534	5	0.2543707	8.1629045	640	0.0127545	19.943544	1.662E-18
V3	0.0427051	5	0.008541	81.653387	640	0.1275834	0.0669446	0.9969234
V4	0.0974868	5	0.0194974	3.5727908	639	0.0055912	3.4871387	0.0040624
V5	2.4927183	5	0.4985437	149.47299	639	0.233917	2.131284	0.0600675
V6	1.9570056	5	0.3914011	71.856209	639	0.112451	3.480636	0.0041173
V7	2.9533661	5	0.5906732	23.156811	640	0.0361825	16.324824	3.554E-15
V8	1.2133368	5	0.2426674	4.9595034	640	0.0077492	31.315052	1.48E-28
V9	2.0282039	5	0.4056408	11.174897	640	0.0174881	23.195243	1.924E-21
V10	2.648728	5	0.5297456	10.543908	640	0.0164749	32.154794	2.839E-29
V11	0.7282199	5	0.145644	8.8286963	640	0.0137948	10.557861	9.314E-10
V12	1.5528336	5	0.3105667	8.3461655	640	0.0130409	23.814852	5.338E-22
V13	2.4312332	5	0.4862466	24.893475	639	0.0389569	12.481648	1.413E-11
V14	1.8827053	5	0.3765411	14.607954	638	0.0228965	16.445369	2.765E-15
V15	0.5754754	5	0.1150951	20.825718	639	0.0325911	3.5314873	0.0037073
V16	0.0697523	5	0.0139505	4.2864573	639	0.0067081	2.0796517	0.0661873
V17	0.3952188	5	0.0790438	259.35019	640	0.4052347	0.1950567	0.9644134
V18	0.0672715	5	0.0134543	5.2502282	640	0.0082035	1.6400727	0.1472515
V19	2.9867582	5	0.5973516	12.405879	639	0.0194145	30.768291	4.399E-28
V20	2.8092265	5	0.5618453	21.149025	636	0.0332532	16.895985	1.064E-15
CZAS	29.594181	5	5.9188362	76.651473	1689	0.0453828	130.42038	0
DMIN	30.159691	5	6.0319382	90.941489	1917	0.0474395	127.15017	0
DMINK	29.602907	5	5.9205814	76.647168	1916	0.0400037	148.00069	0
DV1	12.336987	5	2.4673975	67.55652	1919	0.035204	70.088508	0
DVK	50.229255	5	10.045851	102.11056	1919	0.0532103	188.79524	0
ABS	26.532622	5	5.3065243	100.71791	1917	0.0525393	101.00097	0
EKS	21.83297	5	4.3665939	137.07039	1914	0.0716146	60.973497	0
DVSR	23.515905	5	4.7031811	91.880533	1919	0.0478794	98.22978	0

SS_Effec – sums of deviation squares,
SS_Error – the within group variance,
MS_Effec – getween groups variance,
MS_Error – mean square error,
F – F statistic,
P – significant level.

Annex 3. Classification results obtained with the neural net

Syllables	Classification			Classification neural net			Errors
	N	A	F	N	A	F	
wi	1	0	0	0.999140,	0.000000,	0	
zyj	1	0	0	0.270711,	0.000000,	0.001453	
ne	0	1	0	0.061830,	0.854130,	0.029656	
go	1	0	0	0.996365,	0.000040,	0	
ki	0	1	0	0.078490,	0.852090,	0.031074	
nna	1	0	0	0.995943,	0.000066,	0	
noc	1	0	0	0.815399,	0.178767,	0	
ne	0	0	1	0.231000,	0.000000,	0.969692	
go	0	0	1	0.000000,	0.000000,	1	
zsym	1	0	0	0.996357,	0.000050,	0	
pa	0	1	0	0.117367,	0.798471,	0.003704	
tja	1	0	0	0.995634,	0.000072,	0	
wspo	1	0	0	0.891583,	0.000030,	0	
mi	0	0	1	0.789550,	0.000000,	0	3
nam	0	0	1	0.216625,	0.000000,	0.987578	
czlo	1	0	0	0.997438,	0.000000,	0	
wje	0	0	1	0.799093,	0.000000,	0	3
ka	0	0	1	0.000072,	0.000000,	1	
ktu	0	1	0	0.082234,	0.848338,	0.031757	
ry	1	0	0	0.995934,	0.000067,	0	
sie	1	0	0	0.994960,	0.000103,	0	
zmniej	0	0	1	0.000000,	0.000000,	1	
szal	0	0	1	0.013068,	0.000000,	1	
film	0	1	0	0.000000,	0.996440,	0.001342	
zga	1	0	0	0.995543,	0.000000,	0	
tun	0	1	0	0.000000,	1.000000,	0	
ku	1	0	0	0.996370,	0.000053,	0	
fi	0	1	0	0.082264,	0.849118,	0.031585	
kcji	1	0	0	0.996374,	0.000053,	0	
nna	1	0	0	0.934174,	0.019122,	0	
ko	0	0	1	0.235810,	0.000000,	0.996099	
wej	0	0	1	0.000000,	0.000000,	1	
A	1	0	0	0.996064,	0.000062,	0	
wienc	0	1	0	0.000000,	0.000000,	0.999971	2
ra	0	1	0	0.080455,	0.850007,	0.031443	
czejj	1	0	0	0.966530,	0.000000,	0.00001	
fan	1	0	0	0.869872,	0.086106,	0	
ta	1	0	0	0.814502,	0.179708,	0	
sty	0	0	1	0.426213,	0.502255,	0	3
czny	0	0	1	0.000027,	0.000000,	0.999742	
i	1	0	0	0.996018,	0.000057,	0	
ra	0	1	0	0.000000,	1.000000,	0	
czej	1	0	0	0.991900,	0.000098,	0	
nie	0	1	0	0.002005,	0.844790,	0.009233	
nna	1	0	0	0.995605,	0.000077,	0	
ko	0	0	1	0.228063,	0.000000,	0.999883	
wy	0	0	1	0.000000,	0.000000,	1	
A	0	1	0	0.000000,	1.000000,	0	
le	1	0	0	0.816054,	0.177644,	0	
za	0	1	0	0.000000,	1.000000,	0	

to	1	0	0	0.991332,	0.000115,	0	
po	1	0	0	0.814200,	0.178759,	0	
bu	1	0	0	0.996371,	0.000053,	0	
dza	1	0	0	0.870416,	0.085575,	0	
jon	0	1	0	0.000153,	1.000000,	0	
cy	1	0	0	0.996249,	0.000056,	0	
wy	1	0	0	0.996371,	0.000053,	0	
bra	0	0	1	0.000000,	0.000000,	1	
znie	0	0	1	0.000000,	0.000000,	1	
by	0	1	0	0.000000,	1.000000,	0	
la	1	0	0	0.885677,	0.060984,	0	
to	0	1	0	0.000000,	1.000000,	0	
o	1	0	0	0.995251,	0.000076,	0	
po	0	1	0	0.000000,	1.000000,	0	
wiesc	1	0	0	0.996336,	0.000054,	0	
o	1	0	0	0.996369,	0.000053,	0	
pa	0	0	1	0.225843,	0.000000,	0.999959	
nu	0	0	1	0.000000,	0.007442,	0.977443	
ktu	0	1	0	0.000000,	1.000000,	0	
ry	1	0	0	0.992654,	0.000176,	0	
za	0	1	0	0.025069,	0.976432,	0.008807	
czol	1	0	0	1.000000,	0.000000,	0	
sie	1	0	0	0.996335,	0.000054,	0	
zmniej	0	0	1	0.002757,	0.913692,	0.026527	3
szac	0	0	1	0.006142,	0.000000,	1	
o	1	0	0	0.820854,	0.105778,	0	
czy	1	0	0	0.996349,	0.000054,	0	
wi	0	1	0	0.000000,	0.999396,	0.000272	
scie	1	0	0	0.995597,	0.000078,	0	
po	0	1	0	0.082356,	0.849034,	0.031508	
za	1	0	0	0.995517,	0.000000,	0	
tym	1	0	0	0.995765,	0.000072,	0	
pa	0	0	1	0.225667,	0.000000,	0.999959	
nem	0	0	1	0.070837,	0.000000,	1	
wszy	0	1	0	0.082440,	0.848913,	0.031303	
scy	1	0	0	0.984974,	0.000947,	0	
to	0	1	0	0.083186,	0.847781,	0.02972	
ba	1	0	0	0.993870,	0.000153,	0	
ga	1	0	0	0.800097,	0.000011,	0	
te	1	0	0	0.851289,	0.114261,	0	
li	1	0	0	0.760278,	0.219084,	0	
zo	1	0	0	0.995762,	0.000060,	0	
wa	0	0	1	0.000000,	0.000166,	0.995945	
li	0	0	1	0.065731,	0.000000,	0.999952	
zo	0	1	0	0.048063,	0.382365,	0.15282	
nna	1	0	0	0.984484,	0.001008,	0	
po	1	0	0	0.996368,	0.000053,	0	
wta	1	0	0	0.828701,	0.075962,	0	
rza	0	1	0	0.000000,	0.999578,	0.000082	
la	0	0	1	0.272905,	0.000000,	0.230781	3
mu	0	0	1	0.000000,	0.000073,	0.996183	
ze	0	1	0	0.009394,	0.996849,	0	
jest	0	1	0	0.421409,	0.470125,	0	
wca	0	1	0	0.224295,	0.574549,	0.000059	
le	1	0	0	1.000000,	0.000000,	0	
du	0	0	1	0.062800,	0.000024,	0.998873	
zy	0	0	1	0.000000,	0.000000,	1	

Acknowledgements

The study has been performed within the 8 T11C 02309 research project financed by KBN.

References

[1] N. CAMBELL, *Automatic detection of prosodic boundary in speech*, Speech Communication, 13, 345–354 (1993).

[2] G. DEMENKO, W. JASSEM, M. KRZYŚKO, *Classification of basic F0 patterns using discriminant functions*, Phonetica, **41** (1988).

[3] G. DEMENKO, L. RICHTER, P. SERWAT, *Analiza statystyczna wyników percepcyjnej oceny granic frazowych i akcentu w języku polskim*, Materiały XLIII Seminarium z Akustyki, 167–172, (1996).

[4] G. DEMENKO, *Wyznaczniki fonetyczno-akustyczne granicy frazy i akcentu w języku polskim*, Technologia Mowy i Języka, vol. 1, 101–124, Wrocław 1997.

[5] P. DUMOUCHEL, D. O'SHAUGHNESSY, *Prosody and continuous speech recognition*, Proceedings of the 3rd European Conference on Speech and Technology, Eurospeech93, 2195–2198, Berlin 1993.

[6] L. FRĄCKOWIAK–RICHTER, *The duration of Polish vowels*, Speech Analysis and Synthesis, v.3, 87–115, PWN, Warszawa 1973.

[7] K. HIROSE, A. SAKURA, H. KONNO, *Use of prosodic features in the recognition of continuous Speech*, Proceedings of ICSLP44, Yokohama, 1123–1126, (1994).

[8] J. IMIOŁOCZYK, I. NOWAK, G. DEMENKO, *Implementacja systemu syntezy ciągłej mowy polskiej z tekstu ortograficznego wprowadzonego z klawiatury komputera typu PC*, Prace IPPT, Warszawa, 11, (1993).

[9] W. JASSEM, *Akcent języka polskiego*, PAN, Wrocław 1962.

[10] W. JASSEM, G. DEMENKO, *Fonetyczno-gramatyczna spójność frazy*, Technologia mowy i języka, vol. 1, 125–139, Wrocław 1997.

[11] D. SCOTT, *Duration as a cue to the perception of à phrase boundary*, Journal of Acoustical Society of America, **71**, 996–1007 (1982).

[12] M. STEFFEN–BATÓG, *Analiza struktury przebiegu melodii polskiego języka ogólnego*, Rozprawa doktorska, Poznań 1963.

[13] R. TADEUSIEWICZ, *Sieci neuronowe*, Warszawa 1993.

[14] C. TRABER, *Data-driven prosody generation using automatic learning procedures*, Recueil des Publications et Communications Externes du Departement RCP, CNET, 159–162 (1997).

[15] J. VAISSIERE, *The use of prosodic parameters in automatic speech recogniton*, [in:] Recent Advances in Speech Understanding and Dialog Systems, H. NIEMANN, M. LANG, G. SAGERER [Eds.], Springer Verlag, **46**, 71–99 (1988).

[16] J.M. ZURADA, *Introduction to artifical neural systems*, West Publishing Company, 1992.