Grzegorz Krynicki

# Classification of Isolated Pitch Patterns by Means of Nearest Neighbour Classifier, Neural Networks and Discriminant Function

with Applications in Polish-English Speech Translation, Intonation Teaching and Lexicography

Praca magisterska napisana w
Instytucie Filologii Angielskiej
na Wydziale Neofilologii
Uniwesytetu Adama Mickiewicza w
Poznaniu
pod kierunkiem
prof. Włodzimierza Sobkowiaka

GALILEI *Meine Absicht ist nicht, zu beweisen, daß ich bis-*
*her recht gehabt habe, sondern: herauszufinden, ob. Ich sage:*
*laßt alle Hoffnung fahren, ihr, die ihr in die Beobachtung ein-*
*tretet. Vielleicht sind es Dünste, vielleicht sind es Flecken,*
*aber bevor wir Flecken annehmen, welche uns gelegen kä-*
*men, wollen wir lieber annehmen, daß es Fischschwänze sind.*
*Ja, wir werden alles, alles noch einmal in Frage stellen. Und*
*wir werden nicht mit Siebenmeilenstiefeln vorwärtsgehen,*
*sondern im Schneckentempo. Und was wir heute finden, wer-*
*den wir morgen von der Tafel streichen und erst wieder an-*
*schreiben, wenn wir es noch einmal gefunden haben. Und was*
*wir zu finden wünschen, das werden wir, gefunden, mit be-*
*sonderem Mißtrauen ansehen. Also werden wir an die Beob-*
*achtung der Sonne herangehen mit dem unerbittlichen Ent-*
*schluß, den Stillstand der Erde nachzuweisen! Und erst*
*wenn wir gescheitert sind, vollständig und hoffnungslos ge-*
*schlagen und unsere Wunden leckend, in traurigster Verfas-*
*sung, werden wir zu fragen anfangen, ob wir nicht doch*
*rechtgehabt haben und die Erde sich dreht!*

Leben des Galilei, B.Brecht (l9/6/18-35)

ABSTRACT

This work reports on the influence that prosody of selected Polish ambiguous utterances may have on their interpretation and translation into English. A method for parametric description of the pitch curves coextensive with these utterances is discussed. On the basis of the pitch parameters obtained by the above method, the classification of the utterances is performed with respect to their interpretation and translation into English. Three approaches to the problem of classification are presented: Pattern Matching approach in the form of the Dynamic Time Warping algorithm and Nearest Neighbourhood decision rule, the statistical approach based on statistical Discriminant Functions and neural approach in the form of two types of artificial Neural Network: a single-cell perceptron and a feed-forward neural network with one hidden layer. The Nearest Neighbour classifier provides 66.7 – 79.5% correct classification rate depending on the disambiguated word, Discriminant Analysis performance ranged from 82.5 to 95% of correct classifications, and the neural approaches produced correct classifications in 71.2 – 84.9% of cases. The applications of the above findings in Polish-English Spoken Language Translation, Polish-English lexicography and intonation teaching are presented and discussed.

*CONTENTS*

## 0. Introduction

Unlike in languages with lexico-morphological tone accent (e.g. Japanese, Swedish), in Polish the pitch pattern is not functionally and systematically related to the segmental level (Jassem 1983:116). Nevertheless, the role of intonation, which is the function of pitch, is not exclusively expressive either (for a different view see Steffen-Batogowa 1996:163). Differences in pitch may be related to lexical and grammatical differences.

It has been observed that in casual speech some Polish ambiguous words show fairly consistent correlation between their prosodic features and meaning (Krynicki 1999:131). This interdependence is most conspicuous for utterances whose pitch patterns may constitute simple and complete tone units. Among these words are exclamations (e.g. *aha*, *dosyć*), particles (e.g. *akurat*, *tak*) and adverbs (e.g. *dobrze, blisko*). The implementation of this observation in the field of Polish-English Speech Translation and its usefulness for English learners of Polish requires that the choice of these words be additionally restricted by the constraint that at least two of their different senses cannot be rendered by the same English equivalent without "seriously distorting" the meaning of the original. For example, it would be inappropriate to translate the Polish exclamation *aha* as the English phrase *by the way* when it was meant as an expression of understanding, as in the dialogue

$Q^A$ *Jak to się otwiera?*
*How does it open?*

*A* *Bardzo prosto. Lekko przekręcić i podważyć śrubokrętem...*
*Very easy. You turn it slightly and lever it up...*

$Q^A$ *Aha...*
***I see...***

and, conversely, we would not translate *aha* that originally expressed a sudden recollection of something one had almost forgotten as the English *I see...*

$Q^B$ *No to cześć. Przyjdziesz jutro?*
*See you... .Are you coming tomorrow?*

*B* *Chyba nie...* (po chwili) *Aha...chcialem ci cos dać.*
*No, probably not.* (after a while) ***By the way****... I've got something for you.*

A similar phenomenon has also been reported for German discourse particles (Stede & Schmitz, 1999:4) and for English utterance-initial particles (Byron 1997:128).

Additionally, in Polish we find cases of using sentence accent for marking scope and/or structural contrast between potentially ambiguous phrases (the statement-questions in Polish are probably the best known example illustrating this point). Consider two homophonous, except for the prosody, utterances *A* and *B* (capitals mark syllables that carry the nuclear tone in the main clauses; the focus-marked constituent is taken into square brackets):

*Q*$^{A}$ Czy zgodził się na wszystko o co go prosiłeś?

*A* *Z*godził się [NA to], abym zabrał się do pracy.
He agreed to my getting down to work.

*Q*$^{B}$ Dlaczego on się na to zgodził?

*B* On [ZGOdził się] na to, abym zabrał się do pracy.
He agreed in order that I could get down to work.

Analogous constructions have been reported for German (Hunt 1994:169), Spanish, Italian and English (Hirschberg et. al. 1995:I-175). However, the detection of sentence stress and sentence boundaries will not be discussed in this work.

For the purpose of this study 5 polysemous expressions were selected. **It was hypothesised that the strategies Polish native speakers employ to disambiguate these utterances in their speech are mainly and consistently restricted to the modification of fundamental frequency and the temporal arrangement of the pitch curve.** Intensity was not considered for technical reasons. (Its direct influence is marginal in comparison with $F_0$ and temporal arrangement, see the results of the classification § 1.2.3) In order to verify this hypothesis, the words and sentences in question were presented in disambiguating contexts to a group of 40 native speakers of Polish (for the contexts used, see APPENDIX). All the utterances were recorded and their prosodic features constituted the input to the neural and statistical classifiers. It was postulated that if such a classifier can be trained to correctly disambiguate between the senses of the utterances it "hears" on the basis of their $F_0$ and temporal features, the hypothesis outlined above would be substantiated. The support for this hypothesis would be particularly strong if the model obtained in the process of the supervised learning provided correct disambiguation rate on datasets other than these on which the network was trained.

It must be noted here that the necessity of the perception studies as a final test for the above hypothesis is recognised. It is planned to utilise the parameters obtained in the process of training the network for generating pitch patterns that are supposed to be distinctive for the meanings of the words and sentences they accompany. Such material will be tested on Polish

native speakers to check the compatibility of the machine classification and the human judgements. At the present stage of research, however, the model reflects the production strategies only.

## 1. Classification of Isolated Pitch Patterns

Mathematical modelling of a real-world phenomenon consists in a possibly true description of its characteristics by means of a function, equation, Neural Network, etc. The usually infinite complexity of real-life phenomena, through the modelling process, is usually reduced to the amount of detail that is manageable for human or machine processing powers and, at the same time, that allows the greatest possible emulation or simulation of reality. The classification of utterances with respect to their meaning can be viewed as mathematical modelling of human judgements about the interpretation of these utterances. The complexity of human speech has to be reduced to the features that, if submitted to e.g. a learner of the language, can be mastered relatively easily and, if given as input to a computer, can be processed efficiently. In the present study, the idealisation of the human classification mechanism takes the form of a Nearest Neighbour classifier, Neural Network and of a Discriminant Function. The input arguments of the Neural Network and of a Discriminant Function classifiers are the characteristics of the utterances that are relevant for their interpretation and the weights expressing the relative importance of these characteristics for the correct classification. The input arguments of the Nearest Neighbour classifier are the smallest distances between the new utterance and all the utterances from the training set. The values of all of these two-value classification mechanisms, or dichotomisers, are the possible interpretations of the utterance we classify.

The creation of the classifier that would adequately model human judgements about the interpretation of certain utterances had to be preceded by the collection of an empirical material. Utterances obtained from Polish native speakers were collected into a spoken corpus of 400 tokens. For every token, its pitch track was extracted, normalised with respect to its frequency and duration and subjected to the original method of parametrisation. This last stage produced either a set of parameters that constituted input to the Neural Networks classifier and Discriminant Analysis classifier, or it produced a set of optimal distances between new observations and the model utterances.

### 1.1. Experimental material

The collection of the experimental material for training the classifiers proceeded in three stages. Firstly, the utterances of the ambiguous words had to be recorded in a machine-readable form. Secondly, the spoken corpus had to be analysed with respect to the prosodic features of the ambiguous items it contained. Thirdly, the pitch track extracted from the corpus had to be either parameterised (Neural Networks and Discriminant Analysis) or time-normalised (Nearest Neighbour DTW classifier). Pitch parametrisation was to eliminate the details of the pitch that were assumed to be of little consequence to the classification process, that is, to create a model of a simple pitch pattern for Polish. DTW time-normalisation was intended to produce an objective measure of similarity between any two pitch tracks.

#### 1.1.1. Spoken Corpus

All the data for the spoken corpus was obtained through the simulation of the natural dialogues on the basis of the previously prepared typescripts (APPENDIX). The disadvantages of this approach are immediately obvious: unnaturalness of the phonetic material collected in that way may often seriously affect the results, the more so, that the words we are interested in, particularly exclamations, express intense emotions and attitudes, which are difficult to imitate in the artificial situation of the recording session.

However, the collection of the spoken corpus in non-simulated environment would be prohibitively time-consuming or/and expensive. If the frequency of the words in question in natural speech was twice that of their frequency in the written corpus (no data about the frequency of these items in natural speech is known to the author), and if the corpus provided fairly balanced information about each of the two meanings of these words, the collection of the adequate amount of data would require several dozen hours of recording, which would exceed the resources available to the author. For example, in the corpus of 1.2 mln words *akurat* occurs 44 times. Assuming only two meanings are represented by these occurences: *tell me another* and *perfectly*, and each of the two meanings is represented by approximately the same number of occurrences contributed by different speakers, and assuming that the average rate of speech is 180 words per minute, the recording time would be over 50 hours.

##### 1.1.1.1. Typescripts

All the data for the corpus was obtained on the basis of two lists, each containing 5 dialogues. The dialogues provided contexts that disambiguated the words or phrases they contained. In order to assure that the variable of the utterance interpretation is dependent on the

prosodic features of that utterance only, that is, to exclude the interference of the factors that could not be measured, these dialogues were invented in compliance with the following priorities:

- parallel dialogues that illustrated two different pragmatic contexts of a given item were to be as similar as possible in terms of their form: typography, length, who initiated the dialogue (researcher or informant);
- if an item was presented in a sentential context in one dialogue, it had to be presented in a very similar, preferably the same, sentential context in the dialogue that illustrated its other meaning.
- the dialogues were intended to imitate everyday conversations as closely as possible;
- the ambiguous word or phrase could not be the only one contributed by the speaker to a single dialogue. Every dialogue contained at least, in most cases exactly, two lines (entries) that were separated by the researcher's line. This step was taken in order to avoid the excessive attention of the speakers on the ambiguous items.

#### 1.1.1.2. Collection Procedure

40 subjects were asked to participate in an app. 3-min. session each. During the session, 5 short dialogues were conducted with the experimenter on the basis of printed transcripts discussed above. Every dialogue contained an ambiguous word in one of its two senses. It's meaning was determined by a disambiguating context. 20 informants out of 40 read the dialogues from the list presented in APPENDIX where the ambiguous words were presented in one set of disambiguating contexts, and the remaining 20 subjects contributed the dialogues from the list presented in APPENDIX with the same potentially ambiguous words but in different disambiguating contexts. In all, the corpus consisted of 200 (5×20 + 5×20) dialogues illustrating different senses of the ambiguous words.

No speaker read two dialogues with the same potentially ambiguous word. The subjects were not informed of the purpose of the study. The informants were instructed to be natural in their responses and relaxed.

#### 1.1.1.3. Selection of the Ambiguous Items

From the corpus of all dialogues, the individual ambiguous utterances were manually extracted. The extraction of most utterances did not pose any difficulties, as they constituted short, complete and independent tone-units. All the clicks, moments of hesitation, etc. (e.g. in - Dobrze się czujesz? - **Hmm**... dobrze..., see APPENDIX) were not cut out as the training set was to be later applied as the reference set in the classification of real-life recordings (e.g.

collected via the Internet). These, on turn, could not be pre-edited by an untrained user of the classifier nor, at the present stage of research, could they be pre-edited automatically.

### 1.1.2. Pitch Track Preparation

When the simple structured pitch tracks of ambiguous utterances were manually extracted and collected to form a smaller corpus, they were submitted to pitch extraction[1], pitch track parametrisation and formed the training and test sets for the neural and statistical classifiers. The pitch tracks used for training the classifiers were used for testing the model[2].

In the discussion of the algorithm, several notions will be used that deserve their definition at this point. They provide a description for continuous frequency string and specify the conditions that must be met by two patterns that are corresponding.

> **Definition 1**. Frequency function $f$ defined on a closed time interval $[t_i, t_k]$ is continuous on this interval if, given the constant sampling value threshold $h$, for all the available time argument measurements $t_j$ , $t_i \le t_j \le t_k$ , it holds that either $f(t_j) \ge h$ or $f(t_j) < h$, but never both $f(t_j) \ge h$ and $f(t_j) < h$.
>
> In the first case we will call the $(k\text{-}i)$-tuple of all $f(t)$ a ***continuous frequency string,*** or a ***string of contiguous (frequency) data points;*** in the second case this ordered string of all f(t) will be referred to as ***zero string***.

[1] For the time being, the pitch is computed using WinCECIL v2.2 ©Summer Institute of Linguistics 1994-97. All the tone contours were smoothed by the software-internal procedure in order to correct and incorporate data points which were incorrect by one or more octaves and to interpolate across single data point gaps. The parameters adopted for pitch extraction were the same for all contours:

| PITCH EXTRACTION PARAMETER | VALUE |
|---|---|
| voicing threshold | 40 Hz |
| minimal number of contiguous data points per string (frequency values given every 0,05 ms, c.f. t'Hart et al. 1990) | 6 items |
| percentage change of the string | 5% |

[2] The method used for testing the effectiveness of all the classifiers is known as **leave-one-out method**. It consists in:

1. training a classifier on all the training set observations except one observation;
2. the observation that has not been used for training the classifier is used as a test observation; the results of the classification is being recorded;
3. stages 1 and 2 are iterated until all the observations have been used once as the test observation;
4. the results of all the classifications (the number of these results equals the number of observations in the training set) are averaged and given as the final classification result.

**Definition 2.** Let $v_1$ and $v_2$ be any different frequency strings (be it voiced or unvoiced, i.e. zero, ones). Two continuous strings of measurements (voiced or unvoiced)

$x = (x_m, ..., x_{n-1}, x_n)$ and $y = (y_o, ..., y_{p-1}, y_p)$,

$x \subseteq v_1$, $y \subseteq v_2$,

are ***corresponding*** if the sets of the measurements

$w = (w_1, ..., w_{r-1}, w_r)$ and $z = (z_1, ..., z_{s-1}, z_s)$,

$w \subseteq v_1, z \subseteq v_2$, $r = m - 1$, s = o –1,

contain the same number of continuous strings of any kind.

#### 1.1.2.1. Interpolation for the missing data points between continuous frequency string

In spite of the smoothing process, some missing points may interrupt the continuous flow of the frequency points. This happens during the pronunciation of e.g. voiceless consonants. If such gaps occur at the beginning of at the end of an utterance, in most cases they can be safely ignored. If the data are missing in the middle of an the utterance, that is between two continuous frequency strings, they have to be reconstructed since the zero strings cannot be time-aligned by means of the Dynamic Time Warping algorithm, nor can they be analysed statistically (§1.1.2.3). The missing frequency data points were recreated by means of a linear interpolation, which visually corresponds to drawing a straight line over the zero string arguments between the closest non-zero values of two different continuous frequency strings.

#### 1.1.2.2. Frequency-normalisation of the Pitch Track

The method for frequency-normalisation was basic (c.f. Jassem, Demenko, Krzysko 1988:5). On the basis of the sufficient number of frequency measurements contributed by a given speaker (Jassem 1983:173), his or her mean pitch and the voice range were computed for the sake of reducing the speaker-dependent variation in the pitch tracks to be classified. After recalculation of the $F_0$ from normal to logarithmic scale, for all the utterances of a given person the mean pitch of 0 and the range of voice ± standard deviation was adopted.

#### 1.1.2.3. Time Normalisation of the Pitch Track: Dynamic Time Warping

Speech is a time-dependent process. Several utterances of the same word are likely to have different durations. Moreover, utterances of the same word with the same duration will differ in the middle, due to different parts of the words being spoken at different rates. To align two patterns (represented as a sequence of vectors) of different length and to obtain a general

measure of similarity, that is, the global distance between them a time normalisation must be performed.

In the early version of the classifier, a simple probabilistic algorithm was applied for aligning two non-corresponding pitch tracks[3]. And whereas it was not terribly wrong to adopt this approach, it was not terribly right either. Intuitive and simple as it was, the algorithm showed significant deficiencies. When integrated into the classification algorithm, it provided correct disambiguation for only 73% of akurat utterances (compared with e.g. 88.5% in Discriminant Analysis) and showed low computational efficiency. In this section a new approach based on the techniques of Dynamic Programming is presented. The algorithm has been implemented on the basis of (Paulus & Hornegger 1998:435-440, Jassem & Demenko 1989:117). The DTW pattern matching algorithm is more efficient and flexible than the probabilistic algorithm. It classifies correctly 76.8% of akurat utterances (the average of 74.8% correct classifications for all 5 utterances, see APPENDIX), it handles different lexical items (not just *akurat*) and allows more diversified pitch pattern structure (any number of continuous frequency strings).

The pseudocode for this process is (Akaidi 2004: 154):

```
calculate first column (predCol)
for i = 1 to number of data points in the input vector
    curCol[0] = local cost at (i,0) + global cost at (i-1,0)
    for j = 1 to number of data points in the reference vector
        curCol[j] = local cost at (i,j) + minimum of global costs (i-1,j),(i-1,j-1), or (i,j-1)
    predCol = curCol
minimum global cost is the value in curCol[number of data points in the reference vector]
```

In order to construct a classifier, the input pattern is taken and the above process is repeated for each template file (from the training set). The template file which gives the lowest global cost is the meaning class estimate for the input file.

[3] As the number of measurements in the frequency vectors we compare almost always varies, (as does the length of voicing that accompanies the utterance), in the former version of the algorithm the longer vector is shortened by a number of values *r* that equals the difference between the highest indexes of both vectors. *r* values are chosen at random from the vector containing greater number of measurements. The frequency values from the shortened vector are then concatenated to form a continuous frequency string which can already have its values compared to the corresponding values from the previously shorter vector. This procedure allows the subtraction of the values from both vectors and in most cases allows to preserve the structure of the pitch curve.

### 1.1.3. Pitch Track Modelling

It was postulated that the native speaker's ability to prosodically disambiguate between two senses of a word can be modelled as a statistical or neural dichotomiser that maps the feature vectors associated with a pitch track to one of the meaning classes corresponding to the senses of words these pitch tracks are coextensive with. Thus, it was suspected that pitch curves can be associated with features that assume different values depending on the meaning class of the word they accompany. Features of distinct meaning classes should be different and separated from each other. If the classifier does not provide a satisfactory classification, the hypothesis about the feature vectors coming from different statistical populations will not be supported. Depending on the method of classification, the significance of results of classification will be understood in a different way (see §1.2).

The modelling task was given to Nearest Neighbour classifier, two types of neural networks: single-cell feed-forward perceptron and a feed-forward neural net with one hidden layer (FF) and finally to a Discriminant Analysis classifier. The input to Nearest Neighbour classifier constituted a general measures of similarity (minimal global distances between the template and new observations). The input to the statistical and neural classifiers constituted the parameters calculated for pitch patterns from the training set and new observations.

The easiest way to obtain a similarity measure for any two time signals is through the computation of the features' distances along the corresponding time and accumulation. However, depending on the speed of speaking, utterances can be stretched or compressed, and, although they essentially may belong to the same class, this fact may not be detected without time-normalisation: the parametrization of the signal or the appropriate pattern-matching algorithm.

For the purpose of the classification discussed here, both approaches to the comparison of the feature sequences with missing correspondences of single features were applied. In the Nearest Neighbour classifier pattern-matchnig algorithm was applied (see §1.1.2.3). For statistical and neural approaches parametrisation method was developed.

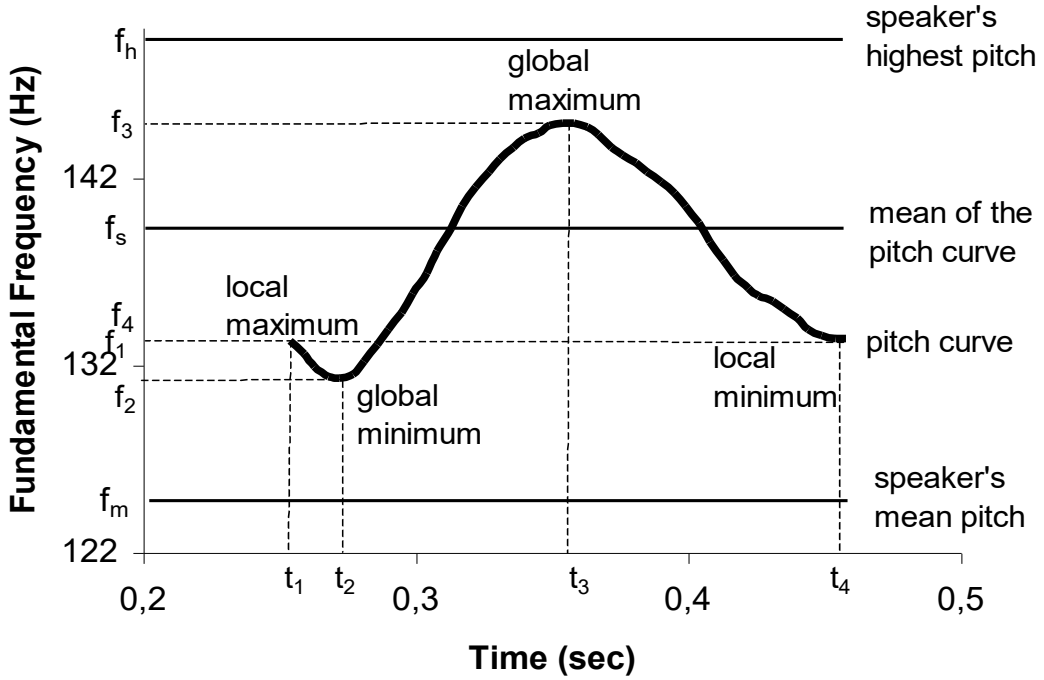
In order to obtain a parametric characteristic of a pitch tracks, it was necessary be establish a possibly small set of prosodic features that were suspected to have any power of discrimination between the pitch curves meant by their speakers to belong either to one or to the other of the two meaning classes. Ideally, all the pitch parameters should be tested for their correlation with the specific readings of the ambiguous word. This, however, if at all possible, would be prohibitively time-consuming. For this reason only the features that intuitively seemed to provide the highest discrimination power were included in the feature space. For the

implementation of the model (see §3, c.f. 1.2.3.3) only those parameters were selected from the feature space that were shown to have the greatest discriminating power.

The extraction of parameters from the pitch curve was performed automatically. The procedure for $F_0$ parametrization operated on a predefined feature space that consisted of the following model parameters:

- global maximum frequency value of the pitch curve relative to the speaker's highest pitch observed in all the utterances he or she contributed to the corpus (cf. Heuft et. al. 1995:379). For the pitch curve in Fig.2, this parameter would be defined as the quotient of the maximal pitch $f_3$ minus $f_m$ and the highest pitch $f_h$ minus $f_m$ (highest pitch is computed as the sum of the speaker's mean pitch and the double standard deviation from this mean, see Jassem 1983:174);
- time argument for the global maximal value of the pitch curve relative to the beginning of the whole pitch curve, that is $t_3 - t_1$ in Fig. 2;
- relative frequency values of the minima that immediately precede and follow the global maximum, which translates into $|f_2 - f_m| / (f_h - f_m)$ and $|f_4 - f_m| / (f_h - f_m)$ below;
- time arguments for the minima next to the global maximal value relative to the beginning of this pitch curve, that is $t_2 - t_1$ and $t_4 - t_1$ in the Fig. 2;
- the number and duration of the strings of contiguous frequency points within every pitch pattern; in the Fig. 2 the number of "continuous" $F_0$ strings is 1, its length is $t_4$ - $t_1$;
- the mean of the pitch curve (unbiased standard deviation); $f_s$ in the picture;

On the basis of these parameters, any the neural and statistical classifiers could be trained.



## 1.2. Different Approaches to Classification: Statistical and Neural Models of Native Speakers' Competence

Trying to decide how to classify a new pitch pattern with respect to the meaning class it belongs to poses many difficulties. Meaning groups defined for most words are not clearly defined. Consider, for example a new data point (the green star in the Fig. 3. below) in the space of three variables extracted on the basis of the utterances of the word *proszę* contributed by all 40 speakers:

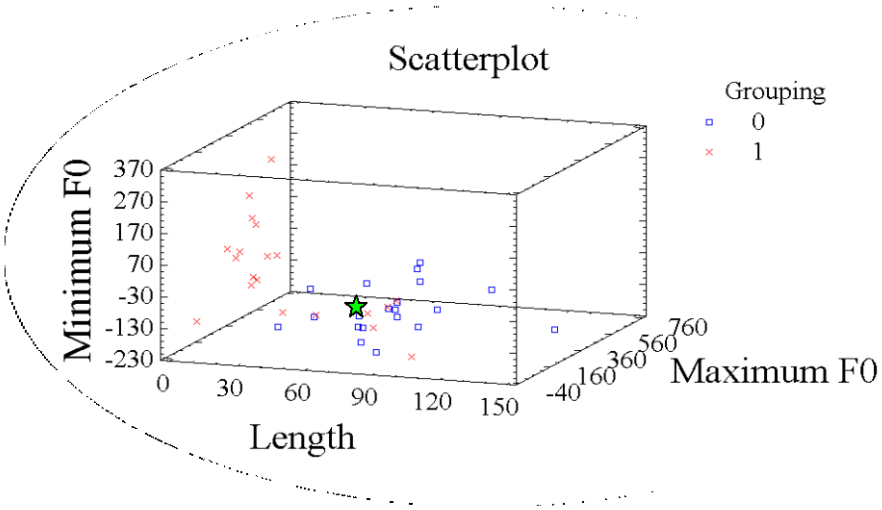



**Fig. 2. Visualisation of the classification problem for two groups, three-dimensional features space and 40 observations. The green star represents new observation of unknown group origin.**

Viewing all three variables simultaneously, it's difficult to visualise exactly how the meaning groups might be delineated. Additionally, we cannot be certain that the variables we have chosen are really the ones that Polish native speakers apply to perform the classifications in their brains. It seems quite reasonable to suspect that, in their ability to discriminate between different meaning groups of the pitch tracks they hear or produce, Poles rely on a range of variables that is beyond the access of machine simulation at the present stage of research on Artificial Intelligence. These variables may not only be prosodic but also pragmatic, semantic, social, etc. However, various mathematical and heuristic tools have been developed that, within a very narrow range of variable space, can model certain aspects of human competence. In this

chapter we shall present some of these tools and compare their effectiveness in modelling intonational competence of Polish native speakers.

Before the modelling procedure took place, empirical data were collected (§1.1.1). On the basis of this corpus, the process of fitting the data into neural and statistical models was undertaken. The aims of the modelling procedure were similar for both neural and statistical approaches to classification:

- to discover whether the speakers employed prosodic means to disambiguate the ambiguous words, i.e. whether the null hypothesis about the feature vectors coming from the same population can be rejected; and if so,
- to find out what characteristics of the pitch track are distinctive for the specific senses of the given word;
- to establish the relative influence of these characteristics on the process of disambiguation;
- to arrive at a two-pattern neural and statistical classifier that would correctly assign new ambiguous utterances to one of the two meaning classes.

Three classification methods were applied in the present study. The simplest method was based on a Dynamic Time Warping algorithm and Nearest Neighbour classifier (c.f. §1.1.2.3). A more sophisticated heuristic method of classification operated on a Neural Network algorithm. Two types of feed forward Neural Nets were tested: Single Cell Neural Network and three layer Neural Network. The last classification method utilised statistical discriminating functions formulated in Discriminant Analysis. Since the collected data do not fully conform to the assumptions of Discriminant Analysis, an additional section (§1.2.3.5) will be devoted to non-parametric statistical significance testing.

### 1.2.1. DTW Nearest Neighbour Classifier

The application of the Dynamic Time Warping algorithm (see §1.1.2.3) for the time alignment and comparison of any two vectors required that they contain only contiguous data points. The continuity between the stretches of contiguous data points of a pitch track (assuming the prior application of the pitch-tracker-internal smoothing function) was obtained through linear interpolation. The missing frequency data points were reconstructed on the basis of their neighbouring values.

The classification task was given to a simple two-value nearest neighbour classifier based on the Dynamic Time Warping (DTW) algorithm.

The input to the nearest neighbour classifier constituted the similarity measures calculated for unknown and reference patterns on the basis of the DTW pattern-matching algorithm (described in §1.1.2.3). The DTW technique allows both the time alignment of frequency measurements and the calculation of the degree of correspondence between them. The similarity between the pitch track of an utterance whose meaning was not known and the pitch track of an utterance whose meaning was already known was expressed by the lowest global distance between all the frequency values of these pitch tracks. The difference between two individual data points in these pitch tracks was expressed as the Euclidean distance.

No separate test set was constructed for the purpose of classification. The pitch track that was to be classified was derived from the training set and compared with all the patterns except itself by DTW routine. Once the minimum distances were computed between the new pitch pattern and all reference patterns, the new pitch track inherited the meaning class of the reference pattern that had the smallest distance from the new pitch track. If the class thus attributed to the new input pattern agreed with the original intention of its author (the information about the origin of pitch tracks was retained through the classification) it was counted as a success. Otherwise, it did not count.

After the classification, the new pattern was returned to the training set and served as a reference pattern for the comparison of other observations.

The correct classification rates ranged from 66.7% for the utterance DOBRZE (see APPENDIX) to 79.5% for the utterances PROSZĘ and NO NO.

| Polish expression | English translations | Correct classification rate of the Polish word with respect to the meaning expressed by the English equivalent | |
|---|---|---|---|
| | | Raw frequency (out of 39 in each case) | Relative frequency (in %) |
| PROSZĘ | Come in!<br>Please, do. | 31 | 79.5 |
| AKURAT | Tell me another!<br>Perfectly! | 30 | 76.9 |
| DOSYĆ | Enough!<br>So so. | 28 | 71.8 |
| NO NO | Well, well!<br>Don't be cheeky! | 31 | 79.5 |
| DOBRZE | All right.<br>Correct. | 26 | 66.7 |
| **Average of correct classification percentages** | | | 74,8 |
| **Standard deviation of correct classification percentages** | | | 5,55 |

**Table 1. Disambiguation rates on the basis of pitch track analysis for five potentially ambiguous Polish words by means of the Nearest Neighbour classifier and DTW algorithm.**

### 1.2.2. Neural Network Classifiers

An alternative approach to classification problem has been implemented in the form of two neural networks. The first network had the architecture of a single-cell perceptron and was trained with a simple weight-incrementation algorithm (a version of a Pocket algorithm, Lee 1994:617). The second network, a more complex but at the same time more effective one, had the form of a Feed-Forward dichotomiser with one hidden layer. The algorithm applied for training this network was the classical backpropagation algorithm (implemented after Żurada et al. 1996:131).

#### 1.2.2.1. Single-cell Perceptron Architecture

The classifier has the form of a single-cell feed-forward neural network (perceptron) with *n* input elements, one hidden layer and a single output. The weighted sum of the input passes thorough a two-value step function that decides on the meaning class the input vector should be assigned to. The network is trained by the pocket algorithm (Lee 1994:617) to obtain the optimal weights assigned to the inputs.

Assume we want classify a new pitch track to one of the classes $\wp$ and $\Re$. Each class contains pitch tracks coextensive with words whose meaning is the same within the class and already known to the system. Every pitch track in these classes is represented as a string of time-frequency pairs or a vector of parameters computed on the basis of this string (see §1.1.3).

Firstly, the differences between the corresponding $n$ parameters (defined in the feature space) of the new pitch pattern and all the reference patterns are computed. From the reference classes, the best matching parameter for the new parameter of unknown class is chosen. The selection is based on a non-optimised nearest neighbour decision rule (Paulus & Hornegger 1998:328, cf. Masters 1996:182): for a new parameter $x_i$, the class that contains a nearest corresponding parameter is selected. The proximity of the two parameters is calculated as the Euclidean distance[4] between them.

Secondly, the minimal global distances between the frequency values of the new pitch pattern and frequency values of every member of both classes are calculated by the DTW algorithm.

Thirdly, the results of the above procedures are committed to a vector and presented as the input to the network.

Thus, given the vector $x = [x_1, x_2, ...,x_{n-2}]$ of $n-2$ parameters computed for a new pitch track $X$ and the vector $q = [q_1, q_2, ..., q_{n-2}]$ of the same number of parameters calculated for the reference pitch track $Q$, and given the minimal global distance between the two pitch patterns $\delta(X,Q)$, the general measure of difference $D$ between the two patterns $X$ and $Q$ can be defined as

$$D(X,Q) = \sum_{i=1}^{n-2}\left(w_i \times d(x_i, q_i)\right) + w_{n-1} + w_n \delta(X,Q)$$

where the vector $w = [w_1, w_2, ..., w_n]$ contains free parameters of the model. These parameters are treated as weights of the neural network and have to be adjusted in the training process to provide the lowest misclassification rate.

Now, if the similarity measure $D$ is computed for the new pitch pattern $X$ and each of the $m$ reference patterns of classes $\wp$ and $\Re$, and for each reference pattern $Q_\lambda$ its meaning class $\zeta(Q_\lambda)$ is known, the classification of the unknown pattern can be done by applying the nearest neighbour classifier (cf. above). For a new pitch curve $X$, the class $\zeta(X)$ is chosen by the decision rule

$$\zeta(X) = \underset{\zeta(Q_\lambda)}{\operatorname{argmin}} \left\{ D(X, Q_\lambda) \mid \lambda = 1,2,...,2m \right\}$$

[4] $d_k(v) = | v - \mu_k |^2$ where $d_k$ is the Euclidean distance between $v$, an observerd measurement vector, and $k$ reference vectors $\mu_k$, c.f. footnote 7 (Schürmann 1996:74)

The difference *D* is then passed as a parameter to the step function that, depending on the value of *D*, decides on the choice of the English equivalent for the utterance whose pitch track constituted the input to the network. In the course of training the network, the interpretations of the inputs performed by the net are compared with intentions of the their human authors and the feedback to the network is given. The training set pitch tracks are presented to the network until all the possible weights are checked. The configuration of weights that provides the greatest number of correct translations is stored in a vector and used in the classification of new cases. This algorithm, known as Pocket Algorithm (Lee 1994:617), is very simple in implementation and most reliable in finding optimal weights, however, it is unacceptable if the network contains a greater number of neurons, input values and connections between the neurons. Checking all the possible weights in the case of a complex network would not be computationally feasible. Thus, in the Feed-Forward Neural Net discussed in the next paragraph a more sophisticated training algorithm has been applied.

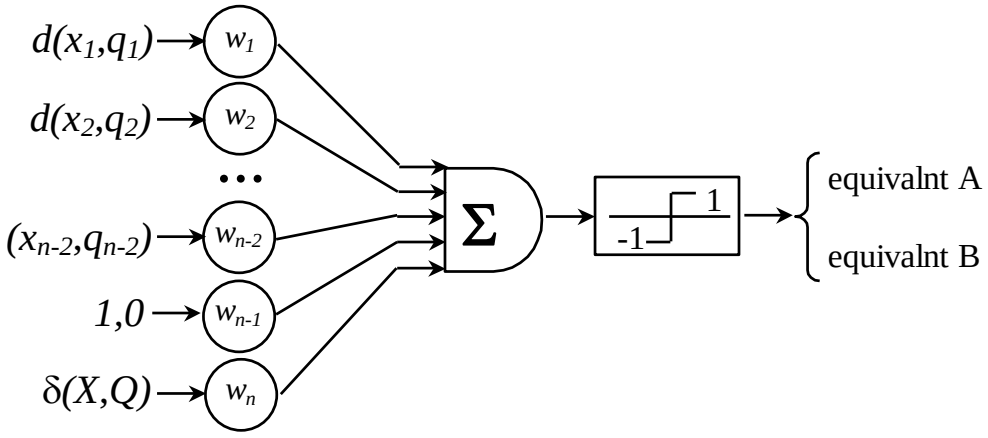



**Fig. 3. A single-cell neural network with *n* inputs: differences between the parameters of the compared pitch tracks, balance input with the constant value of 1,0, and a minimal global distance δ.**

#### 1.2.2.2. Feed-Forward Neural Network with one Hidden Layer and Backpropagation Training Algorithm

The classification task formulated for Feed-Forward Neural Network is the same as in the case of the Perceptron classifier discussed above. The differences are the architecture of both networks, the training algorithm used to optimise the weights and the input values. The detailed description of the algorithm can be found in e.g. Żurada et al. 1996:128ff, Osowski 1996:44ff. Implementation suggestions are given in Masters 1996:94ff.

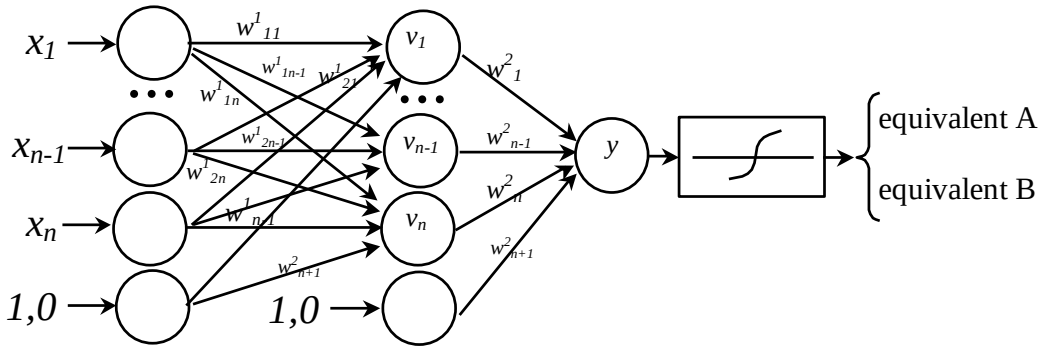



**Fig. 4. A Feed-Forward Neural Network with *n* inputs *x*: the n=7 parameters extracted from pitch tracks and the balance input with the constant value of 1. Sigmoid function decides about the choice of the equivalent. Weights *w* for each layer and connection are optimised in the backpropagation training.**

The results of the classification by means of Feed-Forward Neural Network are given in Table 2 below.

| **Polish expression** | **English translations** | **3-layer Feed Forward Neural Net** |
|---|---|---|
| PROSZĘ | Come in!<br>Please, do. | 71.2 |
| AKURAT | Tell me another!<br>Perfectly! | 84.9 |
| DOSYĆ | Enough!<br>So so. | 84.1 |
| NO NO | Well, well!<br>Don't be cheeky! | 79.0 |
| DOBRZE | All right.<br>Correct. | 82.6 |
| **Average of correct classification percentages** | | 80,4 |
| **Standard deviation of correct classification percentages** | | 5,60 |

**Table 2. Disambiguation rates on the basis of pitch track analysis for five potentially ambiguous Polish words by means of Feed-Forward Neural Network with one hidden layer.**

### 1.2.3. Discriminant Analysis

One of the approaches to the problem of data classification has long been known and applied in statistics. The approach is known as Discriminant Analysis (DA), one of several extensions of Multivariate Analysis of Variance. In one of its variants, it consists in a linear delimitation of the hyperspaces constituted by the features of the classified items. The

coefficients of the functions that define these hyperspaces are adjusted to minimise the misclassification rate.

This method has already been applied in the empirical studies of Polish suprasegmental phonetics by Demenko 1985 and Jassem et al. 1988. The best result achieved in these studies was 85.7% of successful classification for 150 8-dimensional vectors representing $F_0$ parameter measured at 8 equidistant time points for 150 utterances of the word *dobrze*. The classification was performed with respect to eight Polish nuclear tunes obtained from utterances read by untrained informants who imitated the intonation of a phonetician.

The results of both studies cannot be compared since they are of a different level of complexity: 8 groups in the case of Jassem et al. study vs. 2 groups in the case of my study, 150 8-dimensional vectors vs. 40 7-dimensional vectors. The definition of the group differs significantly in both studies. However, the comparison of some methodological aspects of both studies can be attempted. My approach to classification differs from the approach of Demenko and Jassem in the formulation of the data given as the input to the classifier. In the approach proposed by Demenko and Jassem, every the pitch pattern was represented as a vector of 8 frequency values measured at 8 different equidistant times. In the approach adopted in the present study, the parametrization of the pitch pattern was conducted, which was believed to allow a more precise description of the pitch patterns and, in effect, to yield a lower misclassification rate than in the case of frequency sampling. This intuition was tentatively confirmed by a pilot study carried out for the word *proszę*, which was misclassified in 12.5 % of cases under the parameter approach and in 16.5% under non-parametric approach.

Discriminant Analysis for all the speakers, utterances and predictor variables was conducted under STATISTICA ™ 5.0 PL [5]. The Discriminant Analysis classifier has also been implemented as a module of the PAST Classifier[6] in order to enable the user to test the efficiency of this approach to classification without the need of using any statistical package. Nevertheless, for technical reasons, the author was able include only the two most distinctive predictor variables in his implementation, namely *length* and *standard deviation* (see Table 12, §1.2.3.4). In the present chapter assumptions of the Discriminant Analysis will be presented (§1.2.3.1). As Discriminant Analysis is in principle a multi-variate ANOVA, the assumptions of Discriminant Analysis are similar to those of ANOVA. Next, a detailed procedure of finding discriminant variables will be shown. Discriminant Analysis is designed to develop a discriminating function which can help to predict which meaning class a given pitch track

[5] Licence nr. SP7127969005G51

[6] PAST Classifier (Prosodically Aided Speech Translation Classifier) is the implementation of the classification algorithms discussed in the present study. Enclosed on the attached diskette.

belongs to based on the values of the above predictor variables. In §1.2.3.3 the assumptions for DA will be discused. Next, the results of the classification obtained by means of dedicated statistical packages and the author's own software will be presented and analysed (§1.2.3.4). In this analysis, 40 speakers represented cases that were used to develop a model to discriminate among the 2 levels of group: one of two meaning classes for each word. Irrespective of whether the analysis was conducted in STATISTICA or PAST Classifier, always a subset of 7 predictor variables was used. Finally, in §1.2.3.5 several non-parametric tests will be conducted to verify the findings of the DA.

#### 1.2.3.1. Predictor Variables

In the statistical approach the variables used in the process of classification were principally the same as in the neural analysis (§1.2.2). At this stage of research, however, some minor modifications had to be introduced because of the specificity of statistical approach. In the case of Discriminant Analysis we no longer operate on the differences between the model and new observations. For this reason, no pitch track alignment, and hence, no time normalisation is necessary. We also do not need to formulate any general similarity measure external to the classification mechanism as was the case in DTW-based classifier or in the case of Neural Net classifier. In the case of Discriminant Analysis, similarity between pairs of classes in the space of predictor variables is expressed by an analysis-internal Mahalanobis Distance[7].

All 7 predictor variables used in the statistical analysis were:

[7] $d_k(v) = (v - \mu_k)^T \mathbf{C}^{-1} (v - \mu_k)$ where **C** is a covariance matrix between the compared new vector $v$ and the $k$ reference patterns $\mu_k$

| | |
|---|---|
| **VECTOR SIZE** | the number of data points in the pitch track from the first non-zero frequency measurement to the last non-zero frequency measurement inclusive. The zero frequency measurements that separated *continuous frequency strings* (for the definition of *continuous frequency string* see § 1.1.2) were counted in the computation of the size of the vector. However, **neither the actual zero frequency values nor the interpolated values were included in the computation of the pitch track parameters.** |
| **MAX. F0** | maximal frequency value in the normalised pitch track. All the maximal frequency values had to be normalised with respect to standard deviation and multiplied by 100 in order to increase the precision of the results (applies also to the min. F0 parameter). |
| **TIME ARGUMENT OF MAX. F0** | time point at which the normalised frequency string assumes its maximal value. |
| **MIN. F0** | minimal frequency value in the normalised pitch track. |
| **TIME ARGUMENT OF MIN. F0** | time point at which the normalised frequency string assumes its minimal value. |
| **MEAN VALUE OF PITCH TRACK** | mean value of the pitch track produced by an individual speaker. Normalised with respect to the mean pitch of the speaker, as obtained from all the utterances of a given speaker (more than 8000 frequency data points in each case (~200 data points per second), cf. §1.1.2.2, Jassem 1983:173). |
| **STANDARD DEVIATION OF PITCH TRACK** | standard deviation of the pitch track produced by an individual speaker. Normalised with respect to the variance of all the utterances produced by a given speaker. |

**Table 3. Pitch pattern parameters used in Discriminant Analysis.**

### 1.2.3.2. Assumptions of Discriminant Analysis

In order for the Discriminant Analysis to be performed, the data must comply with the following constraints (Jassem 1998:22, Chatfield & Collins1980:125nn, Krzysko 1982:9):

1) The number of classes (groups) distinguished in the analysis must be at least 2. In the case of the present study only two classes are considered. Consequently, the Discriminant Analysis performed here is the most basic one and computationally is analogous to linear regression.

2) Each class must be represented by at least 2 cases. In our case each class is represented by 20 cases.

3) The number of distinct variables must be less than the total number of cases minus the number of classes. We have 7 distinct variables, which is less than the total number of cases (40) minus the number of classes ($7 < 38$).

4) Each of the predictor variables must be measured along an interval or ratio scale. All of the values that represent predictor variables are ratio ones: for all of them multiples and ratios have meaning. The frequency of 123.4 Hz is twice that of as 61.7, the length of a track of 80 data points is twice the length of a track of 40 data points, the time span of 30 ms is twice the span of 15 ms. In the present study, the only variable that contains non-ratio data is the grouping variable, which represents *nominal data*. It assumes one of two values: 0 or 1. These numbers are used as labels for different meaning classes and have no measuring value. However, they are not treated as predictor variable, but rather as classification variable.

5) The within-group variabilities (expressed as within-group correlation matrices) must be approximately equal.

| | LENGTH | TMAX_F0 | MAX_F0 | TMIN_F0 | MIN_F0 | MEAN | STDDEV |
|---|---|---|---|---|---|---|---|
| LENGTH | 1,00 | 0,63 | -0,06 | 0,61 | -0,42 | -0,36 | 0,21 |
| TMAX_F0 | | 1,00 | 0,01 | 0,20 | -0,30 | -0,27 | 0,14 |
| MAX_F0 | | | 1,00 | 0,03 | 0,30 | 0,80 | 0,76 |
| TMIN_F0 | | | | 1,00 | -0,30 | -0,18 | 0,24 |
| MIN_F0 | | | | | 1,00 | 0,71 | -0,31 |
| MEAN | | | | | | 1,00 | 0,34 |
| STDDEV | | | | | | | 1,00 |

**Table 4. The Within-Group Correlation matrix for all utterances (all 40 speakers included).**

The within-class variability for the exemplary word *proszę* shows greater violations of the constraint than the results pulled for all the utterances. Intuitively, the diversity of correlations for all the utterances is not very different from the diversity of correlations observed by Jassem for his data. Correlations in Jassem's study vary from –0.05 to 0.57. It is claimed, however, "that minor violations of this condition are not fatal for the result of Discriminant Analysis". In the following analysis several tables of results will be presented: each for a different configuration of variables depending on how strongly they violate the assumption of homogeneity of variances.

6) None of the random variables must be a linear combination of any other of the random variables. The between-speaker independence of data has been guaranteed by the corpus collection procedure, which made the communication between the informants impossible (see §1.1.1.2). The within-speaker independence of the data was more difficult to assure. The dialogues were read one after another, which, as in the case of all repeated measurements, made the data inherently correlated. As the informants read the dialogues in the order presented in the

APPENDIX, the reader should be more cautious to the results obtained for the last word (*dobrze*) than to the first word (*proszę*). However, in a multidimensional sample, there is, in practice, always some degree of correlation between the random variables, and again, Discriminant Analysis is quite robust to minor violations of this condition.

It is noteworthy to mention that there is strong tendency for the *mean* and the *standard deviation* of a pitch track to be strongly correlated with the *minimal* and *maximal frequency* values of the pitch track. This may be explained by the fact that both minimum and maximum frequency values are included in the calculation of the mean and standard deviation. The lower the minimum value the lower the mean and, usually, the greater the standard deviation. Additionally, a strong correlation appears between the *length* of a pitch track and the *time argument for the maximum frequency* value, as well as between the *length* and the time argument for the *minimum frequency* value.

None of the correlation coefficients is high enough for a variable to be very highly redundant. Additional analysis will, however, be performed for the configuration of variables that does not include relatively highly redundant variables (i.e. *length* and *mean*).

7) each of the classes must have a distribution that is approximately multivariate-normal. The data collected for each predictor variable differ widely in their distributions: *standard deviation* and *time arguments for maximum F0* observed for all utterances across all speakers are exponentially distributed (Fig. 6c, 7a), *maximum frequency* (Fig. 6a) and *mean* (Fig. 6d) have lognormal distribution whereas the *length* (Fig. 6b) density function is roughly normal.

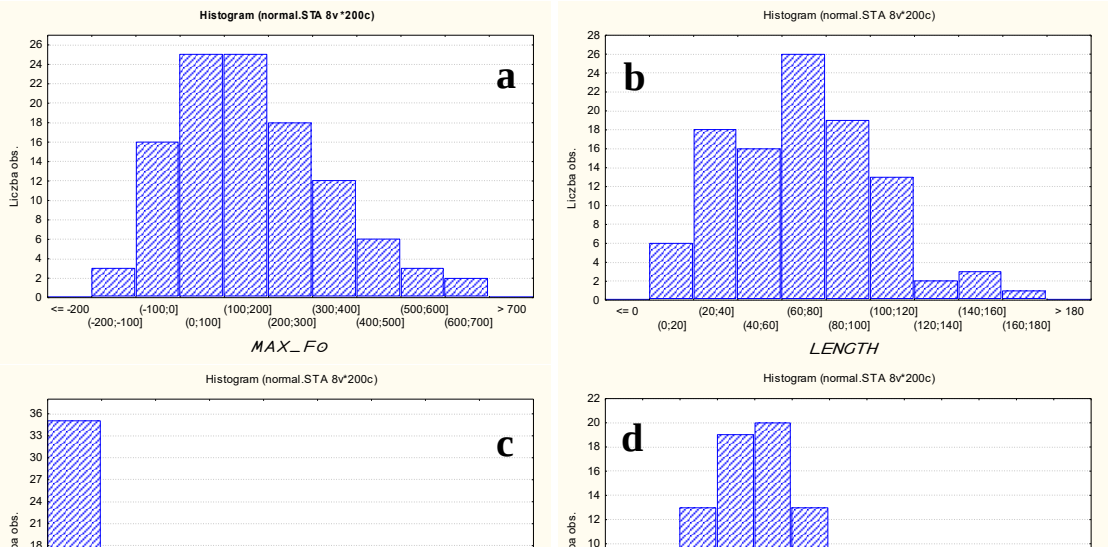



**Fig. 5. Lognormal distribution of *maximum frequency* (a), normal distribution of *length* (b), exponential distribution of *time arguments for maximum frequency* (c) and lognormal distribuion of *mean* variable (d).**

In the present study no adequate solution has been found to overcome this problem. No function could be formulated that would linearly transform all the data so that they conform to normal distribution. We could for example formulate a function for the *length* variable

NUMBER_OF_OBSERVATIONS = 102 * 20 * normal(LENGTH; 72,59803; 39,0466)

that would transform the LENGTH data from lognormal to normal distribution, but the application of similar transformation to e.g. *standard deviation* values would result not in a normal density function. S*tandard deviation* histogram before and after transformation is shown in Fig.7. In principle, the distribution of the variable does not change after the tranformation.

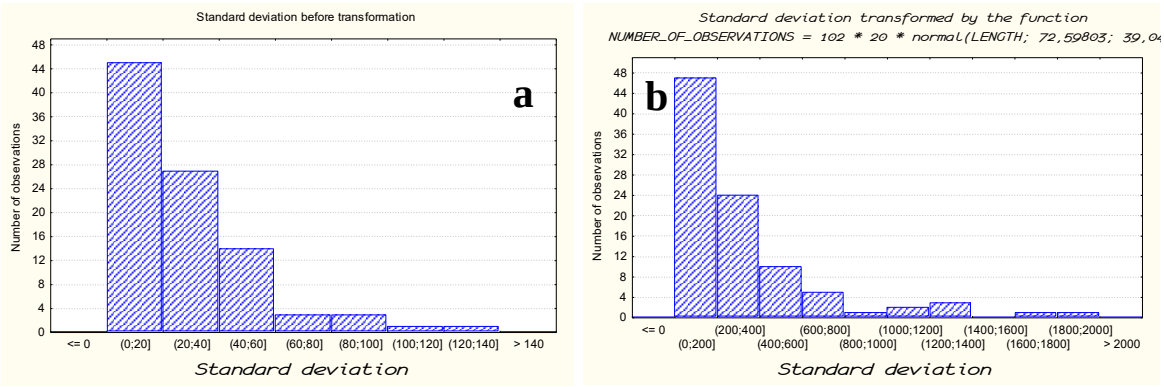



**Fig. 6. *Standard deviation* histogram before (a) and after (b) transformation**

For this reason, additional non-parametric statistical significance testing has been performed, as reported in § 1.2.3.5., in order to corroborate the findings of Discriminant Analysis. Another solution, which, however, has not been explored in the present study, would be to collect a sample of greater size. This would reduce the non-normality of the data and would make the results of the Discriminant Analysis more reliable, whatever the classification results.

### 1.2.3.3. Detailed Methodology

In the following paragraphs, a detailed procedure of Discriminant Analysis approach to classification shall be presented. This material was based on Mathcad™ [8]Help files and STATISTICA™ Manual[9]. The implementation of this algorithm for two variables is included as a module of the PAST classifier.

#### 1.2.3.3.1. Introduction

[8] MathSoft Inc. (1997). Mathcad 7 Professional for Windows [Computer program help].
[9] StatSoft, Inc. (1997). STATISTICA for Windows [Computer program manual].

Having found the numerical representation for every single pitch pattern from the training set (for the process of pitch pattern parametrization, see §1.1.3), the observations can be subjected to classification. The algorithm presented below operates on a two-dimensional feature vector plus one grouping variable. In this form, the algorithm has been implemented in the PAST classifier. Analogous operations are performed on seven-dimensional feature vectors plus one grouping variable by statistical package. The results of the classification of eight dimensional vectors in STATISTICA are discussed in §1.2.3.4.

Consider 40 parametrized pitch patterns of *prosze* defined as 3-dimensional vectors in the matrices Group 0 and Group 1. The values in the first column indicate whether the utterance which the pitch pattern comes from was meant as *Come in!* (0) or *Please, do* (1). The values in the second column represent the length of a pitch track vector. The normalised maximum frequency values for pitch tracks are listed in the third column (see the §1.1.3.).

$$\mathbf{GROUP}_0 := \begin{bmatrix} 0 & 76 & 436.209034 \\ 0 & 92 & 221.340718 \\ 0 & 83 & 350.23272 \\ 0 & 129 & 558.166418 \\ 0 & 115 & 348.871109 \\ 0 & 109 & 81.478212 \\ 0 & 95 & 27.685453 \\ 0 & 89 & 59.084611 \\ 0 & 107 & -9.729193 \\ 0 & 94 & 29.678803 \\ 0 & 81 & -19.734704 \\ 0 & 90 & -38.670885 \\ 0 & 93 & 53.26893 \\ 0 & 81 & -25.236301 \\ 0 & 39 & 67.70984 \\ 0 & 77 & 54.551973 \\ 0 & 83 & -23.25674 \\ 0 & 54 & 56.781415 \\ 0 & 74 & 132.437126 \\ 0 & 61 & -16.289174 \end{bmatrix} \qquad \mathbf{GROUP}_1 := \begin{bmatrix} 1 & 25 & 138.391946 \\ 1 & 81 & 36.296463 \\ 1 & 82 & 153.938423 \\ 1 & 46 & -0.855126 \\ 1 & 97 & 92.583481 \\ 1 & 84 & 180.668207 \\ 1 & 8 & 332.645857 \\ 1 & 18 & 367.656248 \\ 1 & 76 & 147.521689 \\ 1 & 16 & 228.905524 \\ 1 & 14 & 292.470137 \\ 1 & 12 & 216.31022 \\ 1 & 20 & 181.779802 \\ 1 & 46 & 213.911351 \\ 1 & 17 & 319.565757 \\ 1 & 15 & 148.192649 \\ 1 & 14 & 263.323297 \\ 1 & 13 & 407.742982 \\ 1 & 10 & 163.502688 \\ 1 & 9 & -20.62179 \end{bmatrix}$$

Graphing the length and maximum frequency parameters separately, we can see that the groups appear to be only fairly distinct with respect to length and frequency of the maximum.

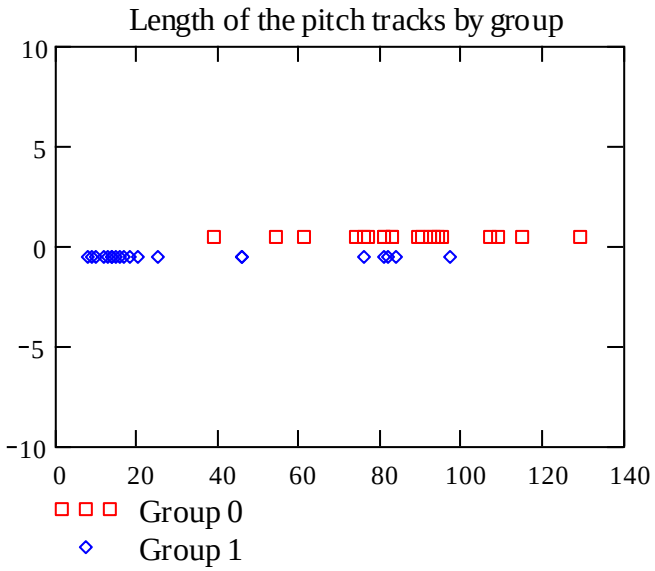



**Fig. 7. Time arguments for maximal frequency values for pitch patterns from both groups. Maxima of proszę (come in) are more dispersed in time than the maxima of proszę (please, do).**

... and normalised maximal frequency in a given of a pitch track.

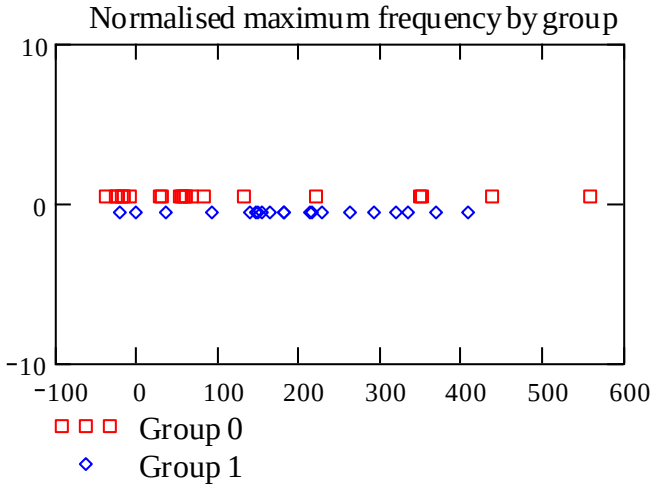



**Fig. 8. Maximum frequency values for pitch patterns from both groups.**

Plotting both variables simultaneously, it is difficult to see how the groups might be distinguished. For instance, would we categorise a new pattern (denoted by a + on the graph) having the time of the maximum 76 (i.e. 76$^{th}$ measurement, each every 0.005 sec) and frequency 216 (i.e. 216 Hz after the normalisation with respect to the mean) as belonging to Group 0 or Group 1?

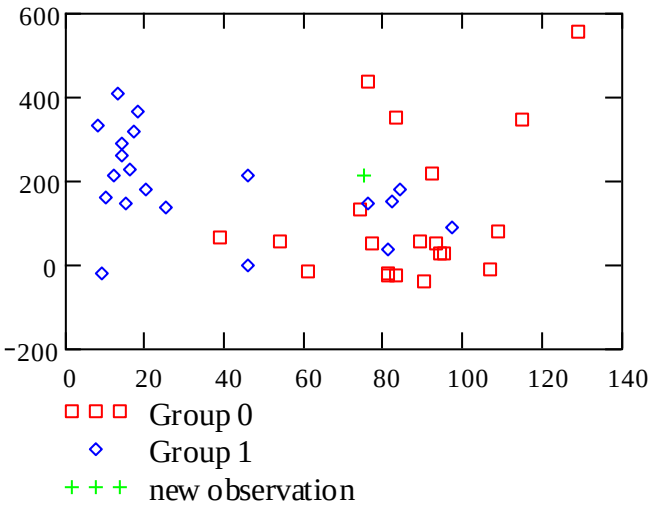



**Fig. 9. The assignment of a new observation is not straightforward given the parameters of 40 pitch patterns. Should the new observation be assigned to the group of its nearest neighbour, i.e. the Group 1? Or should the classification be guided by the fact that a relatively small proportion of Group1-occurrences have the length of their pitch tracks of around 80 data points?**

#### 1.2.3.3.2. Discriminant Function (DF)

In general, what we want is to find a combination, *DF*, of the *p* predictor variables (obtained as parameters of the pitch tracks)

$$DF = \beta_1 \cdot X1 + \beta_2 \cdot X2 + \beta_3 \cdot X3 + \ldots + \beta_p \cdot Xp$$

that will maximise the distinction between groups. In other words, we'll estimate values for β coefficients so that the variation between groups is greater than the variation within the groups. The method is the same as the one used in an analysis of variance for finding differences in group means. Here, though, since we have more than one variable, covariances as well as variances are used.

In the case of two variables, vector length X1 and maximum frequency X2, the DF would have two coefficients $\beta_1$ and $\beta_2$ that maximize the distinction between the two groups. The DF with two parameters would be

$$DF = \beta_1 \cdot X1 + \beta_2 \cdot X2$$

After the application of statistical routine for calculating $\beta_1$ and $\beta_2$ (Fisher coefficients), the following parameter estimates are obtained

$$\beta_1 = 0.1848 \qquad \beta_2 = 1.3494$$

Once we have a set of coefficients, we can form discriminant values for each observation in each group. Discriminant values (DV) are calculated as the sum of products of the Fisher coefficients and variables in both groups. DV calculated for both groups are given below (f = 1..20)

| Group 0: $(\mathbf{DV}_0)_f$ | Group 1: $(\mathbf{DV}_1)_f$ |
|---|---|
| -4.1971 | -1.3965 |
| -6.0388 | -5.8122 |
| -4.9777 | -5.5178 |
| -7.6933 | -3.3678 |
| -7.323 | -6.8068 |
| -7.7194 | -5.5806 |
| -6.8633 | 0.4539 |
| -6.3262 | -0.1682 |
| -7.858 | -5.0989 |
| -6.7839 | -0.4554 |
| -5.9872 | -0.1105 |
| -6.7048 | -0.2021 |
| -6.637 | -0.8952 |
| -6.0044 | -2.6969 |
| -2.6415 | -0.2453 |
| -5.4625 | -0.6344 |
| -6.1445 | -0.2015 |
| -3.773 | 0.3228 |
| -4.9997 | -0.2208 |
| 4.5134 | 0.7228 |

**1.2.3.3.3. Cutoff point**

In order to use the above values to distinguish observations in one group from observations in the other, a cutoff value of some sort is needed.

Below, the discriminant values for each group have been graphed, along with their average locations (red and blue horizontal lines across the Fig. 11), from which we obtain the average locations for both groups (black dashed line in the middle of the graph). We will adopt the centre of the average locations (dashed black line) as a cutoff value that would allow the split of the feature plane into two meaning classes.

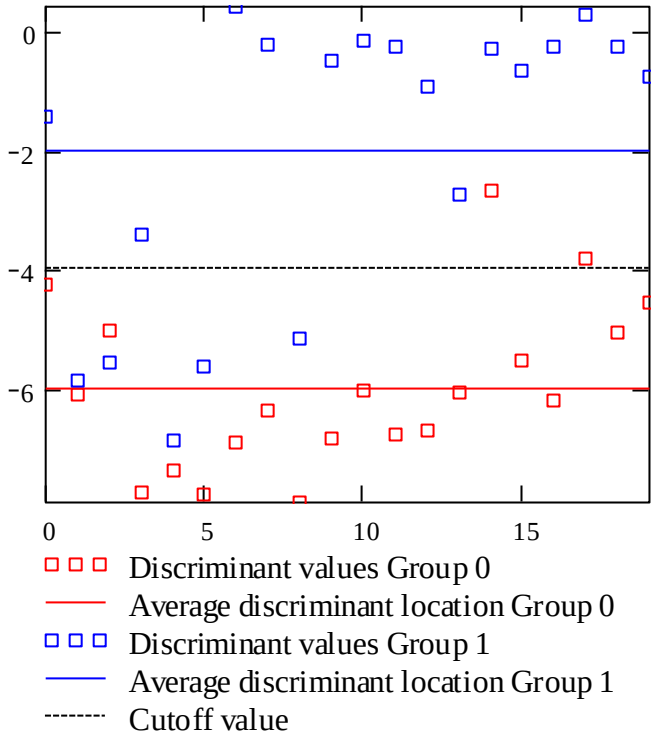



**Fig. 10. Cutoff line (black dashed marker) between the DVs of two groups.**

If an observation has a discriminant value less than Cutoff we'll classify that observation as belonging in Group 0. Otherwise, the observation will belong in Group 1. As can be seen in the Fig. 11, some of the observations will be classified incorrectly.

#### 1.2.3.3.4. Results of the classification

For the purpose of classification, a three-column matrix for each group is created, CLASS0 and CLASS1. The first two columns will list the observation index and the original group memberships. The last column will list group assignments based on the discriminant function.

**CLASS0** =

| | | |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 0 | 0 |
| 2 | 0 | 0 |
| 3 | 0 | 0 |
| 4 | 0 | 0 |
| 5 | 0 | 0 |
| 6 | 0 | 0 |
| 7 | 0 | 0 |
| 8 | 0 | 0 |
| 9 | 0 | 0 |
| 10 | 0 | 0 |
| 11 | 0 | 0 |
| 12 | 0 | 0 |
| 13 | 0 | 0 |
| 14 | 0 | 1 |
| 15 | 0 | 0 |
| 16 | 0 | 0 |
| 17 | 0 | 1 |
| 18 | 0 | 0 |
| 19 | 0 | 0 |

**CLASS1** =

| | | |
|---|---|---|
| 0 | 1 | 1 |
| 1 | 1 | 0 |
| 2 | 1 | 0 |
| 3 | 1 | 1 |
| 4 | 1 | 0 |
| 5 | 1 | 0 |
| 6 | 1 | 1 |
| 7 | 1 | 1 |
| 8 | 1 | 0 |
| 9 | 1 | 1 |
| 10 | 1 | 1 |
| 11 | 1 | 1 |
| 12 | 1 | 1 |
| 13 | 1 | 1 |
| 14 | 1 | 1 |
| 15 | 1 | 1 |
| 16 | 1 | 1 |
| 17 | 1 | 1 |
| 18 | 1 | 1 |
| 19 | 1 | 1 |

By graphing the misclassified observations in Group 0 and in Group 1, along with the original data, we can visualize how the discriminant function separates the observations into distinct meaning groups.

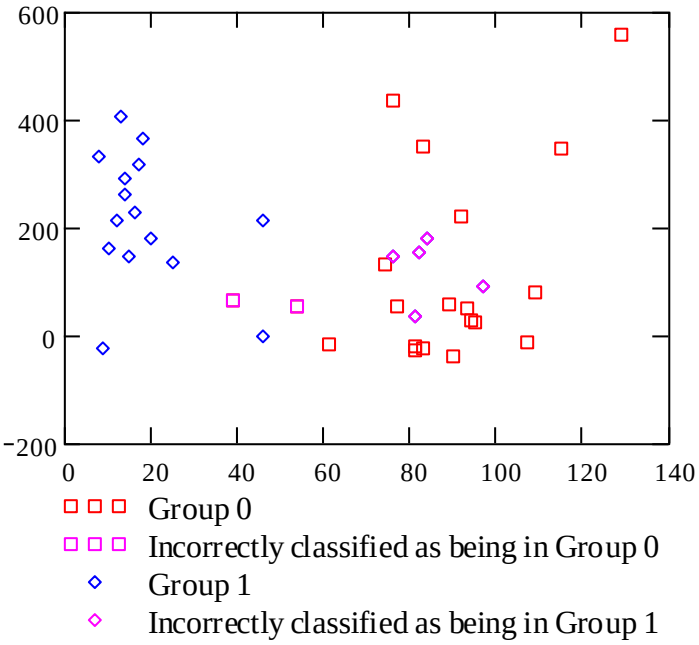



**Fig. 11. The original data plotted against misclassified observations.**

The results of the classification can be summarised in the table of hit and miss rates for each group (Table 5)

| Original group | Total number of cases | Cases classified to Group 0 | Percentages of classified to Group 0 | Cases classified to Group 1 | Percentages of classified to Group 1 | |
|---|---|---|---|---|---|---|
| GROUP 0 | 20 | 18 | 90 | 2 | 10 | |
| GROUP 1 | 20 | 5 | 15 | 15 | 75 | |
| | | | | | | 16.5 = 82.5% |

**Table 5. Classification results for Group 0 and Group 1. The highlighted cells contain correct classifications. 82.5 % of the total observations were classified correctly.**

### 1.2.3.4. Discussion of the Results

The application of STATISTICA allowed the analysis for all of the predictor variables computed for each pitch track. Thus, for example the classification of the *proszę* utterances was successful in 87.5% of cases, compared with 82.5% of cases with the application of only two variables: length of the pitch vector and its maximal frequency value, as exemplified above (§1.2.3.3).

The STATISTICA database had the form of the spreadsheet with rows representing individual speakers and columns representing the parameters calculated for each pitch track. The spreadsheet contained data for 5 different expressions: proszę, akurat, dosyć, no no, dobrze. The *grouping variable* (also known as *classification variable*) indicated the meaning class the utterance was meant to belong to. All the seven *independent variables* (also known as *predictor variable*) were considered.

In the Discriminant Analysis performed here one discriminating function was formulated. for each 2 levels of grouping. As shown in §1.2.3.3. the optimal classification consists in finding the optimal $\beta$ coefficients in the equation of the form

$$DF = \beta_1 \cdot X1 + \beta_2 \cdot X2 + \beta_3 \cdot X3 + ... + \beta_p \cdot Xp$$

where each of the $p$ variables $X_i$ , $0 < i \leq p$, $p = 7$ in our case, is one of the predefined predictor variables extracted from the classified pitch track. The relative magnitudes of the ***classification function coefficients*** are given in Table 6.

| | Group:0 | Group:1 |
|---|---|---|
| | p=,50000 | p=,50000 |
| LENGTH | 0,285594 | 0,127473 |
| T_MAX_F0 | -0,1062 | -0,06957 |
| MAX_F0 | -0,01308 | -0,00578 |
| T_MIN_F0 | -0,06961 | 0,003734 |
| MIN_F0 | -0,03503 | -0,04371 |
| MEAN | 0,169387 | 0,15959 |
| STD_DEV | 0,001336 | -0,01961 |
| Stała | -10,3046 | -4,88488 |

**Table 6. Classification Function coefficients for grouping of the utterance *proszę*.**

Thus, the Classification Function used for the first level of grouping (Group 0) is

$CF_0$ = -10,3046 + 0.285594*LENGTH – 0.1062*T_MAX_F0 – 0.01308*MAX_F0 – 0.06961*T_MIN_F0 - 0.03503*MIN_F0 + 0.169387*MEAN + 0.001336*STD_DEV

and analogously, $CF_1$ for the second level of grouping (Group 1). Each of the *proszę* utterances contributed by 40 speakers was classified by means of these two functions. The variables in that equation (LENGTH, T_MAX_F0, etc) were instantiated by the corresponding parameters extracted for each individual pitch tracks. If $CF_1$ was greater than $CF_0$, the unknown observation was classified as belonging to Group 1. Otherwise, the new observation was classified as coming from Group 0. After summarising all the classification results we obtained the 87.5 correct classification rate for all the observations of *proszę*.

A very important feature of the Discriminant Analysis is its ability to formulate standardised ***discriminant function coefficients***. They provide insights into the linguistic nature of the data and into the psycholinguistic faculties of speakers. From their magnitude conclusions can be drawn about how the individual pitch track parameters are being used to discriminate among the meaning groups. Table 7. shows the standardised coefficients of the discriminant functions used to discriminate amongst the different levels of grouping.

| | |
|---|---|
| LENGTH | 1,831926 |
| TMAX_F0 | -0,50837 |
| MAX_F0 | -0,47303 |
| TMIN_F0 | -0,72158 |
| MIN_F0 | 0,370581 |
| MEAN | 0,178785 |
| STDDEV | 0,203784 |
| Cumulated value | 1 |

**Table 7. Standardised discriminant function coefficients for grouping (*proszę* utterance).**

Thus, the standardised Discriminant Function is formulated as follows

DF = 1,831926*length - 0,50837*t_max_F0 - 0,47303*max_F0 - 0,72158*t_min_F0 + 0,370581*min_F0 + 0,178785*mean + 0,203784*std_dev

DF values obtained by means of this linear equation allow the formulation of different discriminant variables ('roots') depending on the configuration of the pitch parameters included as variables of this equation. Let's perform a stepwise Discriminant Analysis with a backward selection of the variables, which will show **the relative importance of the pitch track parameters for the disambiguation of the** proszę **utterances**. In the backward, stepwise Discriminant Analysis, the variables are removed from the model one by one from the ones of the lowest discriminating power to the ones of the greatest discriminating power. In Table 8 the first column contains the information about how many variables have been removed at that step of analysis. In the first step all the variables are included (0 removed), which also means the discriminability of the Discriminant Function DF is as high as the procedure allows (87.5 %).

| Step, nr of removed variables | name of the variable removed | p-level | nr of variables present | Wilks' Lambda[10] | F | %of correct classific |
|---|---|---|---|---|---|---|
| 0 | | 0,0001 | 7 | 0,42447 | 6,19888 | 87.5 |
| 1 | MEAN | 0,853572 | 6 | 0,424906 | 7,444029 | 87.5 |
| 2 | STDDEV | 0,694206 | 5 | 0,426932 | 9,127604 | 87.5 |
| 3 | MAX_F0 | 0,50254 | 4 | 0,432699 | 11,47191 | 87.5 |
| 4 | MIN_F0 | 0,330375 | 3 | 0,444745 | 14,98177 | 85.0 |
| 5 | TMIN_F0 | 0,077939 | 2 | 0,48542 | 19,61135 | 82.5 |
| 6 | TMAX_F0 | 0,253198 | 1 | 0,503095 | 37,53252 | 82.5 |

**Table 8. Wilks' Lambda for the pitch parameters in the backward, stepwise Discriminant Analysis for the word *proszę***

In the initial step of the analysis (marked as 0), all the 7 pitch parameters are included in the DF function. In this case the result of the classification cannot be worse than at any other step of the analysis. In the first step of the analysis (marked as 1), the parameter of the lowest discriminatory power is removed from the DF equation, namely, MEAN value of the pitch parameter. It should be noted here that the quality of the classification does not suffer from this loss. This is good for sevaral reasons: the MEAN parameter was strongly correlated with other variables for all the words in our analysis, secondly, it was not normally distributed, thirdly, removal of one variable significantly alleviates the computation effort of the classifier. If we know that this variable is unlikely to carry significant discriminatory power, why calculate

[10] A measure of association that is used for describing categorical (nominal or discrete) variables. Lambdas range from 0, when knowing one variable says nothing about another, to 1.0, when knowing one variable always allows to predict the other. Computed as a ratio of within-group covariance matrix total covariance matrix. (StatSoft, Inc. (1997). STATISTICA for Windows [Computer program manual].) See also the discussion of Tables 9-12. p.40

when trying to classify new observations (of proszę!) in the future? Such information is very valuable if we are considering an implementation of a pitch track classifier.

Similar situation appears when we remove *standard deviation* and *maximum F0* parameters. The relatively low discriminatory power of F0 is surprising in the view of findings reported by Heuft et. al. 1995:379 and the literature they cite, that stress the perceptual importance of maximum pitch parameter. One reason for this discrepancy may be the fact that the present study is not so much a perceptive but rather a productive one. Another explanation might be the different grouping variable adopted in the classification performed by Heutft and mine. Heuft classified the collected data with respect to pitch contour groups, not meaning classes as is the case in the present study.

The fact that minimum F0 parameter turns out to have more discriminating power than maximum F0 seems to contradict the findings of Fujisaki 1994:347 about the relative insignificance of minimum F0, especially in comparison with maximum F0. In Heuft's study minimum pitch was not considered at all whereas the maximal pitch was given the highest priority. Nevertheless, if the findings obtained for the single *proszę* word cannot easily, if at all, be generalised to other ambiguous utterances collected in our corpus (see Tables 9-12) , they cannot be extended to the populations they represent and to human pitch perception as a whole.

The time arguments of the minimum and maximum value have approximately the same discriminatory power. A variable that turns out to carry the greatest discriminatory power for *proszę* utterances is the size of the frequency vector, that is the length of the voicing that accompanies the *proszę* utterances. In the dialogues that formed the basis for data collection, *proszę* in the sense of *come in* was significantly longer than *proszę* in the sense of *please do*.

Similar analysis was conducted for the remaining 4 words. The results are tabulated below (Table 9-12). Together with percentages of correctly classified cases, Wilks' Lambda is given for the individual pitch track parameters. In each case the Wilks' Lambda is significant at the $\alpha = 0.0001$ level. The successive values of Lambda increase, which means that as the step analysis proceeds, the pooled discriminability of the remaining variables decreases. $\Lambda = 1$ means that there is no difference between the mean vectors, that is, that the posited classes are indistinguishable on the given variables. A small $\Lambda$ signals good discriminability (Jassem 1998:34).

| | AKURAT | |
|---|---|---|
| | Lambda | % |
| T_MAX_F0 | 0,23 | 97,50 |
| LENGTH | 0,23 | 97,50 |
| MIN_F0 | 0,23 | 97,50 |
| T_MIN_F0 | 0,27 | 97,50 |
| MAX_F0 | 0,40 | 85,00 |
| STD_DEV | 0,47 | 80,00 |
| MEAN | 1,00 | 80,00 |

**Table 9. Wilks' Lambda and percentages of correctly classified pitch patterns by the pitch parameters for *akurat* utterances.**

| | DOSYĆ | |
|---|---|---|
| | Lambda | % |
| MAX_F0 | 0,62 | 82,50 |
| T_MIN_F0 | 0,63 | 82,50 |
| MIN_F0 | 0,71 | 72,50 |
| STD_DEV | 0,74 | 67,50 |
| LENGTH | 0,78 | 67,50 |
| T_MAX_F0 | 0,87 | 67,50 |
| MEAN | 1,00 | 65,00 |

**Table 10. Wilks' Lambda and percentages of correctly classified pitch patterns by the pitch parameters for** dosyć **utterances.**

| | NO NO | |
|---|---|---|
| | Lambda | % |
| MIN_F0 | 0,45 | 82,50 |
| T_MAX_F0 | 0,45 | 82,50 |
| MEAN | 0,47 | 82,50 |
| STD_DEV | 0,47 | 82,50 |
| T_MIN_F0 | 0,50 | 82,50 |
| LENGTH | 0,61 | 80,00 |
| MAX_F0 | 1,00 | 80,00 |

**Table 11. Wilks' Lambda and percentages of correctly classified pitch patterns by the pitch parameters for *no no* utterances.**

| | DOBRZE | |
|---|---|---|
| | Lambda | % |
| MIN_F0 | 0,39 | 92,50 |
| MEAN | 0,39 | 92,50 |
| MAX_F0 | 0,39 | 92,50 |
| T_MIN_F0 | 0,43 | 90,00 |
| LENGTH | 0,47 | 90,00 |
| T_MAX_F0 | 0,74 | 85,00 |
| STD_DEV | 1,00 | 75,00 |

**Table 12. Wilks' Lambda and percentages of correctly classified pitch patterns by the pitch parameters for *dobrze* utterances.**

From the above tables a general discriminant power of all the 7 parameters may be computed. Let's attribute a relative discriminability score to each parameter for individual words. The relative discriminability score will be equal to the number of steps 1..7 in the backward stepwise DA. We obtain the following results

| | proszę | akurat | dosyć | no no | dobrze | total relative discriminability score |
|---|---|---|---|---|---|---|
| MIN_F0 | 4 | 3 | 3 | 1 | 1 | 12 |
| MEAN | 1 | 7 | 7 | 3 | 2 | 20 |
| MAX_F0 | 3 | 5 | 1 | 7 | 3 | 19 |
| T_MIN_F0 | 5 | 4 | 2 | 5 | 4 | 20 |
| LENGTH | 7 | 2 | 5 | 6 | 5 | 25 |
| T_MAX_F0 | 6 | 1 | 6 | 2 | 6 | 21 |
| STD_DEV | 2 | 6 | 4 | 4 | 7 | 23 |

**Table 13. General discriminant power of all the 7 parameters. The higher the total relative discriminability score, the more discriminating power a parameters has.**

From the above calculation we obtain a general measure of discriminability for different parameters. The *length* of the pitch track has the greatest discriminating power (25 total relative discriminability scores). The least important for the disambiguation of the utterances we analysed is the *miminum frequency* of the pitch track (12 total relative discriminability scores).

### 1.2.3.5. Correction of the DA Classification Results for Normality Criterion

Because of the concern about the non-normal distribution of the data, several non-parametric tests were applied to corroborate some of the findings of the Discriminant Analysis. Table 14 shows the results of the application of Mann-Whitney U test for two meanings of ambiguous utterances in the case when only one of the pitch parameters is considered. It should be borne in mind, however, that the comparison of all the parameters separately cannot produce a result that could automatically be set against the results of Discriminant Analysis. In Mann-Whitney U test (Neter 1985:638-41) we cannot compare more than one variable at a time, whereas Discriminant Analysis allows for multivariate comparisons. No non-parametric multivariate analysis is known to the author.

The Mann-Whitney U is a nonparametric test for testing the null hypothesis that the medians of the compared meaning groups are the same. The data from both groups is first combined and ranked from the smallest to the largest. The average rank in then computed for each observation from each group. Then mean of the ranks for each sample is calculated and the assumption that the medians are the same is tested. If p-value is less than 0.05, there is a statistically significant difference amongst the medians of the groups at the 95% confidence level. The only condition for Mann-Whitney U test is that the samples of utteraces from both

meaning groups are randomly selected from each of two populations. This condition has been satisfied, as discussed for Discriminant Analsis in §1.2.3.2.

| | PROSZĘ | | AKURAT | | DOSYĆ | | NO NO | | SKĄD | |
|---|---|---|---|---|---|---|---|---|---|---|
| | U | p | U | p | U | p | U | p | U | p |
| LENGTH | 47,5 | 0,00004 | 177,0 | 0,53385 | 199,0 | 0,97842 | 200,0 | 1,00000 | 120,5 | 0,03152 |
| TMAX_F0 | 112,0 | 0,01730 | 162,0 | 0,30400 | 172,5 | 0,45696 | 124,5 | 0,04113 | 79,0 | 0,00107 |
| MAX_F0 | 118,0 | 0,02655 | 46,0 | 0,00003 | 133,0 | 0,06994 | 68,0 | 0,00036 | 106,0 | 0,01100 |
| TMIN_F0 | 79,5 | 0,00112 | 158,5 | 0,26162 | 131,5 | 0,06390 | 146,0 | 0,14411 | 157,0 | 0,24478 |
| MIN_F0 | 136,0 | 0,08342 | 114,0 | 0,02001 | 130,0 | 0,05830 | 143,0 | 0,12312 | 99,0 | 0,00630 |
| MEAN | 115,0 | 0,02150 | 29,0 | 0,00000 | 118,0 | 0,02655 | 100,0 | 0,00683 | 193,0 | 0,84982 |
| STDDEV | 143,0 | 0,12312 | 70,0 | 0,00044 | 177,0 | 0,53385 | 73,0 | 0,00059 | 80,0 | 0,00117 |

**Table 14. Mann-Whitney U test-statistics and p-values obtained from 35 two-sample comparisons of 7 pitch track parameters (rows) extracted from 5 ambiguous utterances (column pairs). In each of 35 tests, the two levels of the grouping variable were the meanings of the utterance (e.g. *come in* and *please do* for *proszę*), the dependent variable was the pitch track parameter. Differences that are not significant at the 5% level of significance are highlighted.**

From the proportion of non-significant results of comparison it rougly follows that that the least reliable results of DA classification are those for the *dosyć* utterances. The most reliable results are those for the word *proszę* and *skąd*.

The exact reliability measure for the results of DA classification can be obtained by calculating the departure of data from normality distribution. For this purpose several test have been run to determine whether, and if so, how exactly, the pitch parameter data for each word can be modeled by a normal distribution. The applied test were (Neter 1985:8,624) Chi-Square Goodness-of-fit test, the Shapiro-Wilks test, test for standardised skewness and the standardised kurtosis test. Chi-Square Goodness-of-fit test divides the range of the data into several equally probable classes and compares the number of observations in each class to the number expected. The Shapiro-Wilks test is based on comparing the quantiles of the fitted normal distribution to the quantiles of our data. The standardised skewness test looks for the absence of symmetry in the data with reference to the mean. The standardised kurtosis test looks for distributional shape which is either flatter or more peaked than the normal distribution.

The results of these tests for all the parameters of *proszę* utterances are given below:

| | | LENGTH | | TMAX_F0 | | MAX_F0 | | TMIN_F0 | | MIN_F0 | | MEAN | | STDDEV | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Chi2 | Ch2 | 12.50 | 38.50 | 39.80 | 150.30 | 25.50 | 9.90 | 15.10 | 39.80 | 9.90 | 6.00 | 16.40 | 9.90 | 19.00 | 34.60 |
| | p | 0.25299 | 0.00003 | 0.00001 | 0.00000 | 0.00447 | 0.44932 | 0.12846 | 0.00001 | 0.44931 | 0.81526 | 0.08874 | 0.44931 | 0.04026 | 0.00014 |
| Shapiro | W | 0.98 | 0.76 | 0.82 | 0.46 | 0.79 | 0.97 | 0.92 | 0.73 | 0.92 | 0.96 | 0.95 | 0.89 | 0.91 | 0.57 |
| | p | 0.84278 | 0.00016 | 0.00128 | 6.40E-9 | 0.00042 | 0.8364 | 0.12615 | 0.00008 | 0.14102 | 0.65974 | 0.46377 | 0.03619 | 0.08637 | 1.60E-7 |
| Skew ness | Z | 0.30 | 1.34 | 1.66 | 2.63 | 1.83 | 0.06 | 0.94 | 2.04 | 1.31 | 0.98 | 0.89 | 1.40 | 0.91 | 2.81 |
| | p | 0.76447 | 0.17789 | 0.09552 | 0.00837 | 0.06654 | 0.95099 | 0.34288 | 0.04117 | 0.18918 | 0.32482 | 0.37126 | 0.16143 | 0.35892 | 0.00485 |
| Kurtosis | Z | 0.92 | 0.76 | 1.14 | 2.62 | 1.15 | -0.19 | 1.59 | 1.74 | 0.83 | 0.70 | 0.59 | 0.77 | -0.61 | 3.09 |
| | p | 0.35513 | 0.44868 | 0.25196 | 0.00880 | 0.24964 | 0.84199 | 0.11083 | 0.08059 | 0.40614 | 0.47844 | 0.55289 | 0.43769 | 0.54258 | 0.00199 |

marks the p that is lowest among the p's obtained in all tests in the case when this p equals or is greater then 0.10. The parameter (length, etc) for the meaning group (0 or 1) that includes such p-value may be claimed to be normally distributed at 10% level of significance (i.e. with 90% or higher confidence);

marks the lowest p-value in the case when p is less than 0,01 and therefore we can reject the hypothesis that data comes from a normal distribution (with 99% confidence)

**Table 15. Results of 4 normality tests on the pitch tracks of the *proszę* utterances. The tests applied were Chi-Square Goodness-of-fit test (Chi2), the Shapiro-Wilks test (Shapiro), test for standardised skewness (Skewness) and the standardised kurtosis test (Kurtosis). The black border has been drawn around the values of p that indicate normal distribution of both meaning groups.**

Similar tests have been conducted for the remaining 4 words. The abridged table of the results of this analyses is presented below.

| | AKURAT | | DOSYĆ | | NO NO | | DOBRZE | |
|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| LENGTH | 0,00008 | 0,003329 | 0,00065 | 0,00612 | 0,34795 | 0,07604 | 0,12297 | 0,18231 |
| TMAX_F0 | 1,56E-09 | 0,02573 | 9,80E-08 | 2,00E-06 | 2,53E-01 | 1,33E-02 | 6,51E-08 | 6,03E-02 |
| MAX_F0 | 0,000008 | 0,46398 | 0,34921 | 0,144301 | 0,18231 | 0,00045 | 0,19309 | 0,00915 |
| TMIN_F0 | 0,00045 | 0,00459 | 0,297977 | 0,012817 | 0,00447 | 0,08632 | 0,01369 | 0,011121 |
| MIN_F0 | 0,05131 | 0,00151 | 0,00364 | 0,18231 | 0,0003 | 0,00008 | 0,103296 | 0,003328 |
| MEAN | 0,03619 | 0,1754 | 0,18231 | 0,21637 | 0,00569 | 0,000836 | 0,56793 | 0,04026 |
| STDDEV | 1,69E-07 | 0,09542 | 0,012832 | 0,00106 | 0,02344 | 0,000132 | 0,02638 | 0,182311 |

**Table 16. Most conservative results of the 4 normality tests applied to the pitch tracks of 4 utterances:** akurat, dosyć, no no **and** dobrze**, given without the indication which test produced a given result. Minimum p-values are given. Different shades of gray as well as black border have the same meaning as in Table 15.**

Table 17 shows which of the parameters of pitch tracks of which words can ***formally*** be used in any parametric test, including Discriminant Analysis. For the utterances *akurat* and

*no no*, no parameter is normally distributed. This means that formally we cannot perfom DA on the data collected for these words at all. It can be seen that DA will formally be correct if conducted for *dosyć* and *dobrze*, provided it is restricted to *maximum frequency* and the *mean* for *dosyć*, and to the *length* of *dobrze*. The results of Discriminant Analysis that takes these analysis into consideration are tabulated below.

| | % correct |
|---|---|
| PROSZĘ | 62,5 |
| DOSYĆ | 67,5 |
| DOBRZE | 60 |

**Table 17. Percentages of correctly classified pitch tracks for the words that contained at least one parameter that conformed to normal distribution.**

The results are modest. On one hand, this may be the cause of

- small number of parameters that could be considered in the analysis (one parameter in the case of *proszę* and *dobrze*, two in the case of *dosyć*);
- low discriminating power of these parameters; as the backward stepwise DA suggests (see §1.2.3.4, though performed as if all the parameters were normally distributed!), *minimum frequency* is the forth most discriminating parameter for *proszę* (Table 8); *maximum frequency* is the least discriminating parameter for *dosyć* , but *mean* has the greatest discriminating power (Table 10), hence the best result of the DA for *dosyć* here; *length* is the third most discriminating parameter for *dobrze* (Table 12);

On the other hand, however, low classification rates may mean that the differences between the meaning groups are not as big as suggested by other analyses.

### 1.2.4. Discussion of the Results of Statistical and Neural Approaches

The breakdown of the classification results for all the approaches presented above is given in Table 18. below.

| Polish expression | English translations | Relative frequency of correctly classified pitch tracks by approach (in %) | | | | |
|---|---|---|---|---|---|---|
| | | DTW Nearest Neighbour Classifier | Discriminant Analysis | | Perceptron NN[11] | 3-layer Feed Forward Neural Net |
| | | | non-norm | norm | | |
| PROSZĘ | Come in!<br>Please, do. | 79.5 | 87.5 | 62.5 | | 71.2 |
| AKURAT | Tell me another!<br>Perfectly! | 76.9 | 95 | | 73 | 84.9 |
| DOSYĆ | Enough!<br>So so. | 71.8 | 82.5 | 67.5 | | 84.1 |
| NO NO | Well, well!<br>Don't be cheeky! | 79.5 | 82.5 | | | 79.0 |
| DOBRZE | All right.<br>Correct. | 66.7 | 92.5 | 60.0 | | 82.6 |
| **Average of correct classification percentages** | | 74.8 | 88.5 | 63.3 | | 80.4 |
| **Standard deviation of correct classification percentages** | | 5.55 | 6.5 | 3.8 | | 5.60 |

**Table 18. Efficiency of different approaches to the classification of pitch patterns for individual Polish expressions. Discriminant Analysis *norm* means that the classification has been conducted for normaly distributed parameters only. *Non-norm* Discriminant Analysis has been carried out for all the parametrs.**

Dynamic Time Warping algorithm, though the simplest and easiest in implementation, generally results in a higher misclassification rate than any other approach for most words. The only exceptions are *proszę* and *no no*. Pitch tracks of these two words were even less effectively classified to one of the meaning classes (symbolised by English Translations in Table 18 above) by 3-layer Neural Network.

On average, however, 3-layer Neural Network comes out more successful in disambiguating the analysed Polish expressions. Common practice suggest, however, that the 3-layer Neural Network classifier should produce much better results than is the case in the present study. A probable explanation of the poor performance of the NN classifier are: too small training set (for neural nets the optimal sample size is difficult to determine), inadequacies in the training procedure (frequently reported for the classic version of backpropagation algorithm).

The single-cell neural dichotomiser was tested for *akurat* utterances only. The result of this classification approach is least optimistic. Single-cell classifier is simply a neural implementation of liner regression classification (see Fig. 13 below), which is least powerful of all the classification methods. Moreover, this result was obtained for the corpus of 20 utterances only with a less sophisticated parametrization method (see §1.2.2.1).

[11] In the case of Perceptron Neural Net classifier the classification was performed for *akurat* utterance only on the training set half the size used for training the remaining classifiers.

In the above analysis, the best classification results were obtained by means of statistical discriminant functions in the case when all the variables were considered. Intuitively, the higher variability of the classification results obtained in *non-norm* Discriminant Analysis, as indicated by the standard deviation in Table 18, should not undermine the claim that *non-norm* Discriminant Analysis turned out to be the most effective tool for the classification of the sample data from our corpus. A serious flaw of this analysis is that at this stage of research no adequate measure of reliability of these findings can be given since the results do not take into account the departure of the data from normality. An advantage of DA approach (as discussed in §1.2.3.4) is that the discriminant values obtained in the discriminant analysis provide insights into the linguistic importance of the predictor variables for the disambiguation of the Polish words.

Simplified graphical representation of the classification methods discussed above given in Fig.13 (cf. Demenko 1999:148). The known representatives of the two meaning classes are represented by the circle and triangle points in a two dimensional feature space (say, the length of a pitch track and the minimal F0 value found in that pitch track).

The new observation of unknown class is represented by the star.

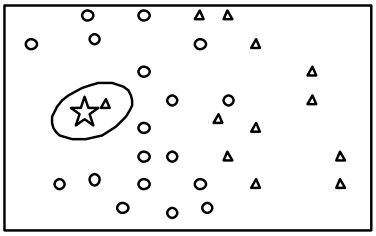
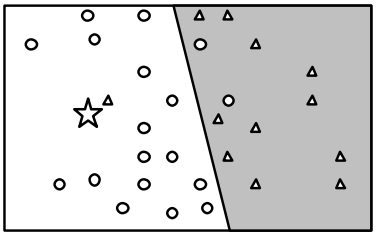
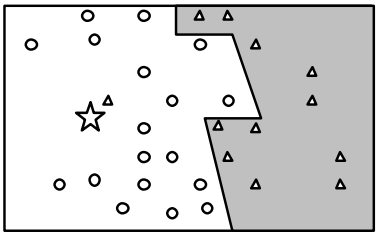
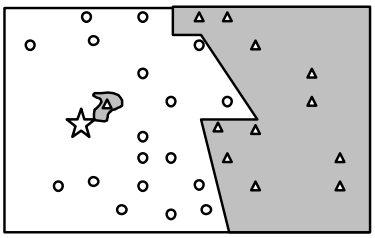
| TYPE OF CLASSIFIER | DISCRIMINATORY POWER OF THE CLASSIFIER |
|---|---|
| Nearest Neighbourhood |  |
| Single Cell Neural Network |  |
| Statistical Discriminant Analysis |  |
| Three-layer Neural Network |  |

**Fig. 12. Visualisation of different approaches to classification in a two dimensional feature space. New pitch track is represented by a star. Training set observations from the two meaning classes are represented by circles and triangles.**

In the Nearest Neighbour classifier, a new observation is assigned to the meaning class of its nearest neighbour. As seen in the picture, even though the new observation is situated deep in the region of the *circle* group, one accidental occurrence of triangle group may decide on the incorrect class assignment of the new observation.

In the single cell neural dichotomiser we are dealing with a simple linear regression problem. Any new observation "to the right" of the cut-off line is classified as belonging to the triangle group, whereas a new observation "to the left" of the cut-off line is recognised as a circle.

In our case, both classification methods are comparable as to their effectiveness.

Statistical discriminant analysis in its linear variant applied here is also based on a linear discriminant function. Discriminant function, however, does not treat all the variables equally, which can roughly be represented as a broken line that separates the two dimensional feature space visualised above.

Theoretically, sufficiently large Neural Classifier can conform to any arrangement of the data points in the defined feature space. There are, however, numerous technical problems: the problem of overfitting the data, which results in the exceptionally low misclassification rate for the training set, but at the same time in equally exceptional high misclassification rate for any new observations; the problem of adequate training procedures that will effectively avoid local minima of the weights (corresponding to the discriminating variables, of 'roots' in the DA approach) assigned to values at respective layers of the network, etc. The discussion of these problems goes beyond the scope of the present study and, which is reflected in the relatively poor results of the neural classification, also beyond the competence of the author.

It should be noted, however, that the relative effectiveness of the classification methods, as suggested by the above table, is only intuitive and should not be generalised to the whole populations of DTW, NN and discriminant classifiers: the discrepancies between the classification results may be merely a result of a sampling error. A longer testing period will be necessary to statistically test the significance of the differences between the classification results provided by each method. [12]

[12] By June 2000, an internet translation service will be made available on-line, which is supposed to allow the collection of a sufficient number of classification results for statistical analysis. The user will be able to access server application through a usual web browser, record one of the ambiguous utterances an hear the server's translation.

# 2. Methodological Aspects of the Study

## 2.1. Methodological Concerns

Apart from the technical problems that were already reported (non-normality of the data, too small a sample, elicited data instead of natural speech), there are methodological concerns about the data and the approach to its analysis.

Firstly, the productive material obtained by elicitation is used for modelling receptive skills. If we wanted to claim that e.g. Neural Network is a model of some aspect of the native speakers' prosodic competence we would face a problem: if the Neural Net is a model of human perception, why has it been taught on the productive material; if the Neural Net is model of human production, why to use it as a perception mechanism. Thus, an assumption has to be made that the strategies people apply to prosodically disambiguate between the senses of ambiguous words **in their speech** are the same as the strategies people apply to disambiguate between the senses of ambiguous words **as they hear them spoken** by somebody else. Such claim, however, has had no scientific justification.

Secondly, only two senses of any ambiguous utterance were considered, when there are sometimes more than two possible. In the case of isolated occurrences of the words listed in the first column of Table 19 the following meanings-translations can be given (on the basis of Szymczak 1978, Stanisławski 1969)

| | |
|---|---|
| PROSZĘ | *Come in!*<br>*Please, do.*<br>*Please.*<br>*What?*<br>*Here you are.* |
| AKURAT | *Tell me another!*<br>*Perfectly!* |
| DOSYĆ | *Enough!*<br>*So so.* |
| NO NO | *Well, well!*<br>*Don't be cheeky!*<br>*There, now!* |
| DOBRZE | *All right.*<br>*Correct.* |

**Table 19. All the main sense-meanings of 5 isolated Polish utterances. The highlighted cells indicate Polish words that have more than two equivalents.**

In the modelling of the data, only two senses were considered. The reason behind it was the significantly higher complexity of the classification of more than two groups, higher risk of wrong classification in such a case and the need for greater amount of data if we wanted to carry out such analysis. This narrowing of the scope of the research may significantly limit the

full applicability of its results in Spoken Language Translation. Before the application of the classification data obtained for the highlighted words in Table 19, additional research will have to be conducted that would provide for additional meanings of these words.

Thirdly, the pitch track characterstics that constituted the input to the classifiers was to some extent intuitive. This means that some important pitch characteristics might have been omited (e.g. the dynamics of the utterance) and some pitch characteristics might have been unnecessarily included in the study (e.g. *minimum frequency*, see Table 13).

Finally, some methodological concerns may be inspired by the nature of the Feed-Forward Neural Network with one hidden layer. The classification results obtained by a multi-layer NN, even if satisfactory, cannot be proven or explained (Masters 1993:87-90). Moreover, multi-layer NN does not provide any theoretical insight into the linguistic nature of the classification, as is the case in Discriminant Analysis. We cannot draw conclusions from the optimal weights about the relative importance of the input parameters.

## 2.2. Interdisciplinary Character of the Research

The solution of the pattern matching problem discussed above rests on the methodologies and findings of a wide range of sciences. It involves methods of several formal and empirical sciences and its consequences are not restricted to linguistics only.

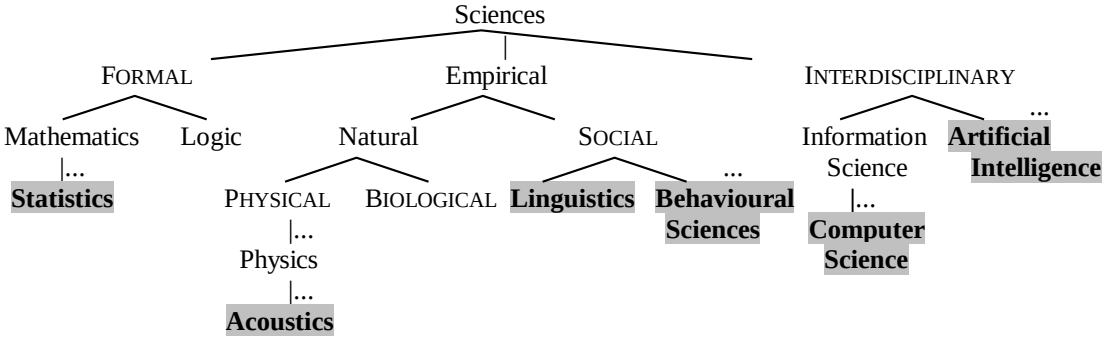



**Fig. 13. Classification tree of sciences. The highlighted leaves represent sciences that are connected with the problem of pitch pattern**

Although most of the claims made in this study have an empirical character, they frequently involved **statistical reasoning** that is independent of the external world. For example, Discriminant Analysis (§1.2.3) is a part of a deductive system based on axioms of mathematical statistics (Krzysko 1982). We also explicitly made several "almost purely formal" statements, as e.g. in §1.2.3.2 where it was claimed that a sample of greater size will

improve the reliability of Discriminant Analysis classification. Such statement can be justified "almost" without referring to the external world:

- on the one hand, Discriminant Analysis assumes normal distribution of the variables for its results to be reliable,
- and on the other, Central Limit Theorem states that if N independent variables (in our case N=40) have finite expected values and variances (here one non-formal assumption: humans cannot produce voice frequencies of infinite range), then their sum for $N \rightarrow \infty$ is normally distributed. In other words, the greater the sample size, the closer will it resemble normal distribution.

Another field that strongly influenced the methodology of this research is **Artificial Intelligence**. Within this paradigm attempts are made to enable computers and machines to mimic human intelligence and sensory processing ability. Within our study two streams of Artificial Intelligence are represented (cf. Winston 1984:2ff):

- traditional statistical approaches that seek to emulate human behaviour without necessarily achieving this end in the way humans do. It is unlikely that any Polish native speaker performs Discriminant Analysis in his brain before s/he decides on the interpretation of some ambiguous utterance on the basis of its pitch track
- and a more contemporary approach that utilises artificial neural networks that are argued to simulate the functioning of the human brain.

Paradoxically, in the study conducted here, the traditional approach gives more insights into our understanding of human intelligence than the neural approaches. It is due to the Discriminant Analysis that we learn about the relative importance of pitch parameters in the human disambiguation of the ambiguous utterances.

The branches of AI research that are represented in our study include machine learning, inference, cognition, knowledge representation, case-based reasoning, natural language understanding, speech recognition and artificial neural networks (Winston 1984:XII-XVIII).

A key technique developed in the study of artificial intelligence is to specify a problem as a set of states, some of which are solutions, and then search for solution states. In our case, each of the two pitch classification options creates a new state. If a computer searched the states resulting from all possible configurations of the model and a new

observation, it could identify those that are most appropriate (in the case of Nearest Neighbour classifier the most appropriate are those that are the nearest in terms of DTW global distance, in the case of Discriminant Analysis the most appropriate are those that are the nearest in terms of Mehalanobis distances). In our study we also explicitly applied a method that is characteristic of AI research: heuristics that limit the search space where we expect to find the optimal global distance between two pitch tracks (§ 1.1.2.3 ).

The affiliation of this study with linguistics is not straightforward. In spite of intensive research in the field, the relationship between the acoustic signal and the linguistic units has not been established. The assumption underlying my research was that the formulation of the linguistic rules that would allow disambiguation of the words in question, their formalisation and later implementation is significantly more difficult than the mechanical, statistical approach to the problem. Thus, instead of "manually" finding the set of semantically distinctive pitch contours and using them for the annotation of the corpus, the available data was automatically classified without the assumption of any linguistic theory. The departure from the linguistic approach to the problem of pitch pattern classification is evident in the case of 2-layer Neural Network classifer that not only does not provide any insight into the process of classifiation but also does not inform the researcher about the importance of the input parameters.

# 3.Applications

## 3.1. Intonational Information in Dictionaries for Machine Use: Prosodically Aided Word Sense Disambiguation in Polish-English Speech Translation

The formalisation of Polish prosody in the context of Spoken Language Translation has gained much attention mainly due to the role that intonation and temporal arrangement play in the segmentation of the spoken input into grammatical units. As has been shown above, however, the classification of utterances with respect to their suprasegmental features may also help in the pragmatic and semantic interpretation of these units and, consequently, may play an important role in their rendering into the target language if incorporated into the "knowledge" of the translating system. Insofar as the ambiguous structures and words cannot be translated compositionally by standard semanto-syntactic methods (e.g. Krynicki 1999:117), the regularities of the prosodic features concurrent with them should be formulated as rules in the dictionary.

Obviously, pitch is not the only factor that can provide a clue as to the most appropriate translation of these words. Additional inputs should be included in a unified model for parsing and translating Polish utterances:

- the frequency of particular senses of ambiguous items in spoken language (e.g. *akurat* meaning *tell me another* is rearer than in the sense of *perfectly*);
- the syntactic context; some of the words discussed above have their meanings that can be fully discriminated on the basis of their position in the sentence only (*Akurat wszedł do pokoju* would not be ambiguous with respect to the senses of *akurat* that could be rendered by English *tell me another* or *perfectly*);
- pragmatic, stylistic and extralinguistic information (e.g. *akurat* in the sense of *tell me another* is more likely in informal contexts)

Prosodic information, along with other types of linguistic and extralinguistic data, may be incorporated as a Multi-Engine Spoken Language Translation system (cf. Frederking 1999:61). The schematic architecture of such a system is proposed in Fig.15

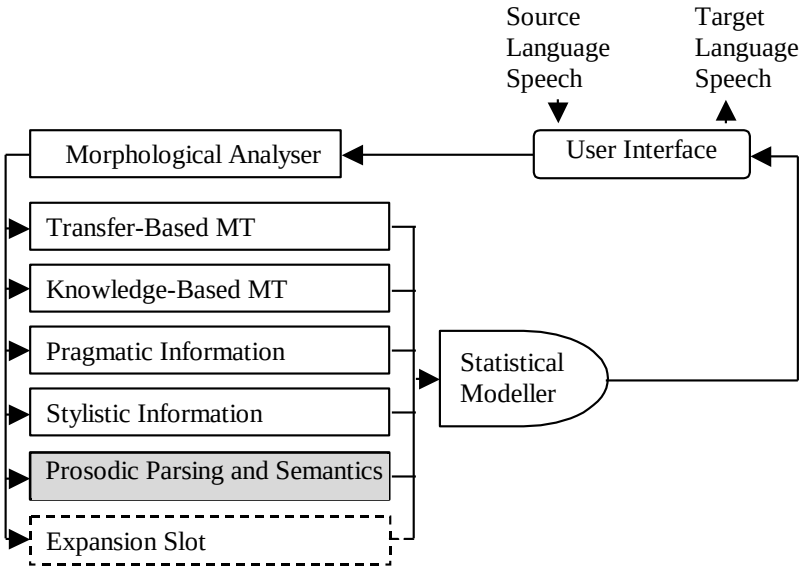



**Fig. 14. Structure of Multi-Engine Spoken Language Translation**

The Multi-Engine SLT architecture makes it possible to utilise all the ranges of information that may play some role in the optimal translation of the source language input. Each translation/optimisation engine attempts to translate/segment the input text. Their products are weighed in the Statistical Modeller with respect to a particular word or context. For example, the weight attributed by the Statistical Modeller to the proposition of the *Prosodic Parsing* module that *akurat* should be translated as *tell me another* should be higher if the *Pragmatic Information* module shows that *akurat* occurs in an informal context.

Context-dependence of prosodic features can probably be detected also for utterances other than just the single-unit particles, adverbs and exclamations discussed above. Such information could be attributed appropriate weights reflecting their real importance for finding equivalents of common lexical words and could be treated on an equal footing with syntactic information.

The **PAST CLASSIFIER** (included on the attached diskette) simulates some features of the Multi-Engine SLT system. It contains a Statistical Modeller, a Prosodic Information module and a Pragmatic Information Module. The implementation operates on a small set of isolated utterances. The user is supposed to inform the machine which word he is going to utter. The disambiguation is performed automatically.

### 3.2. Intonational Information in Dictionaries for Human Use: Applications of Pitch Pattern Classification in Lexicography

The introduction of prosodic information to a Polish-English learner's dictionary seems to be justified by the possibility of the two situations:

- English learner of Polish hears an ambiguous Polish word. He cannot disambiguate this word because e.g. he does not understand the context. However, he can recall the intonation that accompanied its utterance;
- English learner of Polish wants to learn different intonation patterns of an ambiguous word in order to be able to use them in speech.

A small **MULTIPLE ACCESS POLISH-ENGLISH DICTIONARY (MAPED)** has been developed to include prosodic information and allow for complex searches of its database. The prosodic annotation has been provided only for ambiguous headwords of which it was shown or suspected that they can be prosodically disambiguated. In principle, the prosodic transcription adopted in the dictionary conformed in principle to SAMPROSA notation (http://www.phon.ucl.ac.uk/home/sampa/samprosa.htm).

| | Movement | | | IPA | JASSEM | SAMPROSA | MAPED |
|---|---|---|---|---|---|---|---|
| | initial | intermediate | extension | | | | |
| 1 | low | | mid | □ | □ | LM | 01 |
| 2 | mid | | high | □ | □ | MH | 12 |
| 3 | low | | high | □ | □□ | LH | 02 |
| 4 | mid | | low | □ | * | ML | 10 |
| 5 | high | | mid | □ | ` | HM | 21 |
| 6 | high | | low | □ | `` | HL | 20 |
| 7 | mid | (mid) | (mid) | □ | > | M= | 11 |
| 8 | low | mid | low | | □ | LML | 010 |
| 9 | low | high | low | | ^` | LHL | 020 |
| 10 | mid | high | low | | ^` | MHL | 120 |
| 11 | mid | low | mid | | □□ | MLM | 101 |
| 12 | high | mid | high | | `' | HMH | 212 |
| 13 | mid | low | high | | □□, | MLH | 102 |

**Table 20. Different types of prosodic transcription: IPA, Jassem 1999:38, Samprosa and transcription of MAPED. Rows include 13 different categories of pitch patterns.**

The only change that was introduced with respect to SAMPROSA notation was the conversion of the character representations of the three levels of pitch (L,M,H) into numeric

values (0,1,2). Other notations were not considered as they were inconvenient for automatic analysis or incomplete.

In the next stage, all of the pitch tracks in the corpus (described in §1.1.1.) were classified into one of the above pitch curves. The labels of these pitch curves, in turn, served as a basis for prosodic annotation of the dictionary. An ambiguous word whose senses could be prosodically distinguished was assigned the labels of all the pitch curves that accompanied this word in a given sense in the corpus.

MAPED is a database application. The database engine, together with the program, is enclosed on the attached diskette. The application works in Win95/98 environment. It allows for the edition of the existing entries and adding new ones. The program enables conversion of the dictionary into the following formats: RTF, SGML, DOS text, Database Quick Report. It has a built-in SQL query engine that allows complex searches of all the fields of the database.

### 3.3. Intonational Information in Dictionaries for Foreign Learners of Polish

The **ENGLISH INTONATION TEACHER** is a demo application that has been designed to help the learner of English to practice four basic patterns of English intonation. The teaching process consists in repeating the presentation-imitation-feedback cycle until the user's imitation of the presented model is accepted by the program. The software allows some elements of authoring on the part of the user. New model utterances can be recorded without the need for recompilation. For the program to be maximally effective, the model utterances must represent clear cases of the four intonation patterns admitted by the program. The software operates in the Windows '95/'98 environment. The main limitations of the application are the following: the restricted choice of the intonation patterns covered, a very small range of utterances illustrating the use of these patterns, and a relatively poor effectiveness of the classification algorithm for longer utterances (more than ~1 sec). The flexibility and robustness of the program will be the main priorities in its future development.

In the present version of the program the four basic pitch patterns available are: fall, fall-rise, rise, and rise-fall. The simplicity of this model limits the usefulness of the program and is not without adverse effects on the appropriateness of the feedback given to the learner. The user is assumed to utter exactly one of these four patterns and nothing else, which may result in a false acceptance of the user's incorrect patterns.

One of the features of the Intonation Teacher is its modularisation. The program is modelled on the behaviourist teaching process scheme: every stage of the teaching process has its corresponding module in the structure of the program.

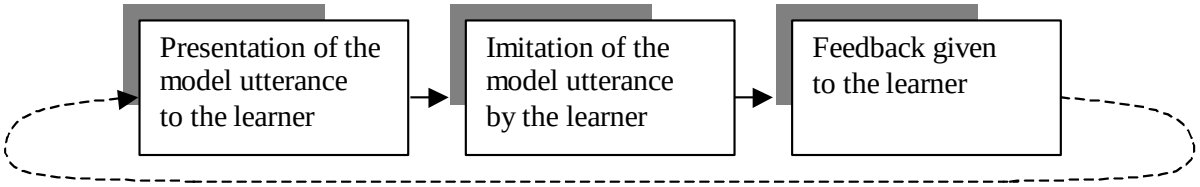



**Fig. 15. The scheme of the teaching process and the corresponding modularisation of the program. The arrows represent the data flow between the modules. The presentation-imitation-feedback cycle is repeated until the learner's utterance is accepted as correct.**

In the presentation phase, the Intonation Teacher allows the model recordings to be chosen and replayed as many times as judged necessary by the learner. In the imitation phase, the user attempts to record his own imitation of the model. At the feedback stage, the suprasegmental features of the user's utterance are extracted and compared with the suprasegmental features of the model by means of DTW algorithm and Nearest Neighbour decision rule (§1.1.2.3). If the least different pattern belongs to the same contour group as the user's utterance, the user's utterance is assumed to be correct and the teaching cycle is stopped. Whatever the judgement of the program as to the correctness of the user's intonation, each time a visual feedback is given to the user in the form of pitch curve set against the model pitch curve.

The program provides several sets of utterances that consist of intonational variants of the same word or phrase. Additionally, there is a possibility of recording new model utterances, provided their pitch curves can be classified as one of the admitted by the program.

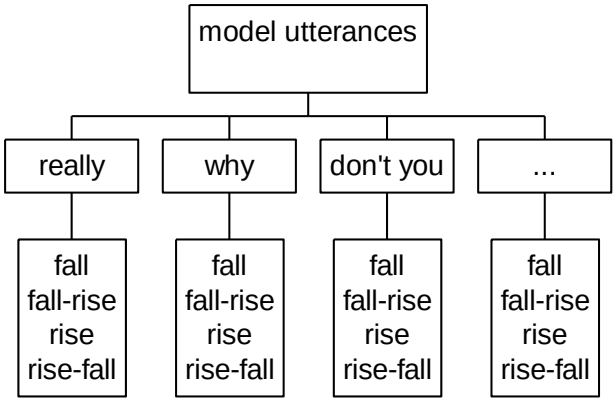



**Fig. 16. The structure of the model corpus. Node with points denotes the possibility of recording user's own utterances.**

Before the teaching process can begin, the user is asked to record a 30 second utterance in order to find his mean pitch (cf. §1.1.2.3). This stage is necessary for the user's pitch track normalisation.

The approach proposed in the Intonation Teacher has many limitations. As already mentioned, the classification algorithm is limited to just a few pitch patterns that are illustrated by a small number of short utterances. Utterances that conform to one of the four basic pitch patterns cannot effectively simulate the diversity of real-life intonation.

What is probably more important, the correct classification rate is around 70% which means that almost one feedback in three is incorrect.

In the future versions of the program a special stress will be put on the optimisation of the classification algorithm, extension of the length of the utterances admitted as models and the speaker independence of the classification.

## 4. Conclusions

In this study the problem of pitch pattern classification has been presented from differenent points of view. Three approaches have been analysed and their results have been compared: Dynamic Time Warping Nearest Neighbour classification, Statistical Discriminant Analysis and the classification by means of Neural Networks. If we disregard violations of data normality, the highest classification rate is obtained by Discriminant Analysis (the average of 88.5% of correctly classified observations). In such a case the least effective algorithm is the DTW classifier (the average of 74.8%). If the classification is conducted within the normality constraint, Discriminant Analysis turns out to be the worst classifier (average: 63.3%), whereas the Neural Net with one hidden layer performs best (average: 80.4%).

In the course of Discriminant Analysis it has been shown that the pitch parameter of the greatest discriminating power is the *length* of a pitch track and that the parameter of the least discriminating power is *minimum frequency*.

In general, the statistical and neural classification results seem to support the main hypothesis of the work, namely, that there exists a fairly systematic relationship between the interpretation of some Polish ambiguous words and their pitch patterns. However, the need for more research in the field has been recognised to strengthen or weaken this hypothesis. The classification of the pitch patterns of these words may find its application in Spoken Language Translation, language teaching and lexicgraphy.

## APPENDIX

### Typescript Group I

| | |
|---|---|
| | (słychać pukanie do drzwi…) |
| **A:** | **Proszę!** |
| | (drzwi się otwierają) |
| B: | Czy można? Mam sprawę… |
| **A:** | **Oczywiście… siadaj… Co się stało?** |
| B : | Wiesz, że Piotr jest w Stanach? |
| **A :** | **Akurat!** |
| B : | Nie wierzysz? Dostałem od niego kartkę… |
| **A :** | **Jak zdobyłby tyle pieniędzy na samolot?** |
| B: | Dobrze się już czujesz? |
| **A:** | **Dosyć.** |
| B: | Nie wyglądasz najlepiej… |
| **A:** | **To tylko zmęczenie.** |
| B: | Skończyłem. Chcesz zobaczyć? |
| **A:** | **Jasne!** |
| | (ogląda portret) |
| B: | Musi jeszcze wyschnąć. |
| **A:** | (Z podziwem kręcąc głową) |
| | **No no…** |
| B: | Podoba ci się? |
| **A:** | **Jeszcze jak! Masz świetną kreskę.** |
| B: | Jak ci idzie praca? |
| **A:** | **Dobrze.** |
| B: | Skończysz do jutra? |
| **A:** | **Tego nie powiedziałem.** |

### Typescript Group II

| | |
|---|---|
| | (otwierają się drzwi) |
| B: | Można? Mam sprawę… |
| **A:** | **Proszę.** |
| B: | Chodzi o to podanie… |
| **A:** | **Tak, wiem… Właśnie rozmawiałem z dyrektorem. Niestety, chyba nie będziemy mogli ci pomóc.** |
| B: | O… przymierzałaś moją koszulę. Jak leży? |
| **A:** | **AKURAT!** |
| B: | Chcesz ją zatrzymać? |
| **A:** | **MÓWISZ POWAŻNIE? BARDZO BYM CHCIAŁA!** |
| B: | Mamo, mogę wyjść na dwóór…? |
| **A:** | **NIE WIDZISZ JAK LEJE!?** |
| B: | Mówiłaś, że jak skończę będę mógł wyjść! |
| **A:** | **Nie marudź. Jak przestanie padać to wyjdziesz.** |
| B: | Ale mamo… |
| **A:** | **Dosyć.** |
| B: | Dam za niego 50 złotych i ani grosza więcej. |
| **A:** | **ALEŻ ON JEST WART TRZY RAZY TYLE!** |
| B: | Nic nie jest wart! Biorę go bo spodobała mi się rama! |
| **A:** | (Wymachując pięścią) |
| | **No no…** |
| A: | Takiś pan groźny? To znajdź sobie innego kupca na te bohomazy! |
| **A:** | **Pytanie za sto punktów: ile księżyców ma Wenus?** |
| B: | ANI JEDNEGO, PANIE PROFESORZE! |
| **A:** | **DOBRZE.** |
| B: | Źle, panie profesorze. Nie można tam oglądać zaćmień słońca! |

**REFERENCES**


Akaidi, M.A. 2004. Fractal Speech Processing, Cambridge University Press.

Byron D., Heeman P., (1997) *Discourse Marker Use in Task-Oriented Spoken Dialogue,* in: [no indication of the editor] Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech), Rhodes, cited after Stede & Schmitz, 1999.

Chatfield Ch., Collins A., (1980) *Introduction to Multivariate Analysis*, London: Chapman & Hall.

Demenko G., (1985) *Klasyfikacja przebiegów częstotliwości podstawowej*, Warszawa : IPPT PAN.

Demenko G., (1999) *Analiza cech suprasegmentalnych języka polskiego na potrzeby technologii mowy*, Poznań : Wydawnictwo Naukowe UAM.

Frederking R., (1999) *Interactive Speech Translation in the DIPLOMAT Project*, in: [no indication of the editor] 1999 Speech Translation Summit Proceedings, Barcelona.

Fujisaki H., (1988) *A Note on the Physiological and Physical Basis for the Phrase and Accent Components in the Voice Fundamental Frequency contour*, in: O.Fujimura (ed.) Vocal Physiology, NY : Raven.

Heuft B. et al., (1995) *Parametric Description of $F_0$-Contours in a Prosodic Database*, in: [no indication of the editor] ICPhS 1995 Proceedings, Volume 2, Stockholm, pp. 378-382.

Hirschberg J. et al., (1995) *The Intonational Disambiguation of Potentially Ambiguous Utterances in English, Italian, and Spanish*, in: [no indication of the editor] ICPhS 1995 Proceedings, Stockholm, I-175.

Hunt A., (1994) *A Generalised Model for Utilising Prosodic Information in Continuous Speech Recognition*, in: [no indication of the editor] IEEE 1994 Proceedings, II-169.

Jassem W., (1983) *The Phonology of Modern English*, Warszawa : PWN.

Jassem W., Demenko G., M. Krzyśko (1988) *Klasyfikacja podstawowych wzorców intonacyjnych z zastosowaniem funkcji dyskryminacyjnych*, Warszawa : IPPT PAN.

Jassem W., Demenko G., (1989) *Zależność przebiegu parametu F0 od długości frazy i dźwięczności segmentalnej*, Warszawa : IPPT PAN.

Jassem W., (1999) *English Stress, Accent and Intonation Revisited*, in: Wiktor Jassem, Czesław Basztura, Krzysztof Jassem (eds) Speech and Technology. Volume 3, Poznań : PTFon.

Krynicki G., (1999) *Suggested Improvements in the Linguistic Aspect of the Electronic Polish-English Dictionary*, in: Wiktor Jassem, Czesław Basztura, Krzysztof Jassem (eds) Speech and Language Technology. Volume 3, Poznań: PTFon.

Krzyśko M., (1982) *Analiza Dyskryminacyjna*, Poznań : WN-UAM.

Kuhn J., (1996) *Context Effects on Interpretation and Intonation*, in: D.Gibbon (ed.) Natural Language Processing and Speech Technology.

Lee L.L., (1994) *On two-pattern classification and feature selection using Neural Networks.* In: [no indication of the editor] Proceedings ICASSP 1994, Vol.2, Adelaide, South Australia.

Masters T., (1996) *Sieci Neuronowe w Praktyce*, Warszawa : WNT.

Morgan D., Scofield C., (1991) *Neural Networks and Speech Processing*, Boston: Kluwer Academic Press.

Neter J. et al., (1985) *Applied Linear Statistical Models*, Womewood, Ill: Richard D. Irwin, INC.

Osowski S., (1996) *Sieci Neuronowe w ujęciu algorytmicznym*, Warszawa : WNT.

Owsianny M., (1998) *Zależność między barwą wokaliczną a częstotliwością podstawową w percepcji samogłosek polskich*, in: Wiktor Jassem, Czesław Basztura, Krzysztof Jassem (eds) Speech and Language Technology. Volume 2, Poznań: PTFon.

Paulus D.W.R., Hornegger J., (1998) *Applied pattern recognition: a practical introduction to image and speech processing in C++*, Wiesbaden : Verlag Vieweg.

Portele T. et al., (1995) *Parametric Description of F0-Contours in a Prosodic Database*, in: [no indication of the editor] ICPhS 1995 Proceedings, Stockholm, II-378.

Saloni M., Świdziński Z., (1998) *Składnia współczesnego języka polskiego*, Warszawa : PWN.

Schürmann J., (1996) *Pattern Classification: A Unified View of Statistical and Neural Approaches*, New York : John Wiley & Sons.

Stanisławski J., (1969) *Wielki słownik polsko-angielski*, Warszawa : Wiedza Powszechna.

Stede M., Schmitz B., (1999) *Discourse Particles and Routine formulas in Spoken Language Translation*, in: [no indication of the editor] 1999 Speech Translation Summit Proceedings, Barcelona.

Steffen-Batogowa M., (1996) *Struktura przebiegu melodii polskiego języka ogólnego*, Poznań : Wydawnictwo Sorus.

Szymczak M., (1978) (ed.) *Słownik Języka Polskiego*, Warszawa : PWN.

t'Hart J., Collier R., Cohen A. (1990) *A Perceptual Study of Intonation: an Experimental-Phonetic Approach to Speech Melody*, Cambridge : CUP, cited after Demenko 1999.

Winston P.H., (1984) *Artificial Intelligence*, *Second Edition*, Reading, Mass: Addion-Wesley Publishing Company.

Żurada J., Barski M., Jędruch W. (1996) *Sztuczne Sieci Neuronowe*, Warszawa: PWN.