

ANALYSIS OF SUPRASEGMENTAL FEATURES FOR SPEAKER VERIFICATION

Grazyna Demenko
Institute of Linguistics
Adam Mickiewicz University, Poznan, Poland
lin@amu.edu.pl

ABSTRACT: The objective of the paper is the assessment of suprasegmental speech features in text-independent speaker verification systems. The linguistic material adopted for research included: read story, read dialogue and spontaneous speech. Fifty speakers without any known speech defects were recorded (3 times of different time intervals). For each individual pitch statistics and dynamic suprasegmental parameters were determined. Selected features were tested with the discrimination analysis and neural networks. Different combinations of pitch parameters were found to be significant for the speaker identification.

I. INTRODUCTION

There are numerous fundamental reasons why suprasegmental speech features are significant for speaker verification systems: (1) prosodic structure is not affected by the frequency characteristics of the transmission systems and the level at which the speaker talks, (2) melodic cues of the utterance are not easy to imitate, (3) individual variations constitute one of the main sources of variables that affect the intonation of the utterance.

Despite a great amount of intensive research conducted over the last few decades, it seems that the practical application of suprasegmentals in speech and speaker recognition is still limited. The complexity of the problem at the stages of reproduction, perception and acoustic analysis of the signal, whether the recognition system is the human brain or a computer, results from its inherent qualities. Each of the four basic acoustic features of a speech signal, i.e. pitch, intensity, duration and speech quality, is carrier of a variety of types of linguistic, paralinguistic and non-linguistic information. For example, a difference in pitch may denote certain element of the system: in English the fall from medium to low pitch has a different function than the fall from high to medium pitch (here we deal with linguistically distinctive phenomena). However, exactly the same difference (from the physical point of view i.e. in respect of the changes of the fundamental frequency) may have gender-specific character and in this case the difference is non-linguistic. Two different sounds of vowels may signal phonematic difference but under certain circumstances the same difference may reflect different voices. A vowel which has two different durations may in one language denote strictly linguistic difference, e.g. in Czech and Polish where the duration difference is phonematic, while the analogous difference of this vowel in Polish has an expressive function and therefore paralinguistic. The melodic structure of utterance relies on the whole range of factors (see e.g. Jassern & Demenko, 1986,) the most important of which include speaker's attitude (discourse conditions), thematic accent placement, lexical and/or grammatical stress realization, segmentally conditioned length of the pitch curve, short term effects of emotional states, physiological long personal voice features, pathological long and short term voice features, speech tempo (long and short term changes in tempo), effects of segmental features, style and dialect.

II. SUPRASEGMENTAL ANALYSIS

For the practical implementation of suprasegmentals in speaker recognition systems, it is necessary to solve the following fundamental methodological and technical problems in the area: a) the reliable extraction of suprasegmental parameters – principally the fundamental frequency, b) selection of the suprasegmental features and their quantitative description, c) description of the long term variability in people's voices, d) integration of segmental and suprasegmental features.

The problem of a reliable extraction of F_0 parameter has been in principle solved, and currently greater emphasis is put on the postprocessor of intonation contour. In the case of the necessity of

precise measurements of pitch level variations (e.g. for the purpose of speech pathology analysis) it is indispensable to verify the extraction manually.

The existing studies devoted to the assessment of the usefulness of suprasegmental features in voice recognition focus mainly on intonation analysis within a given phrase as well as statistic evaluation of pitch variations which characterize given voice in samples of speech which lasts several dozens of seconds. Various methodologies have been applied in comparing intonation contours, see for example Karhunen-Loeve's transformations in the study Atal (1972) or DTW dynamic programming (e.g. [Barlow & Wagner 1998, Mathew & Yegnanarayana & Sundar 1999]). However, the identification of individual voice features on the basis of the analysis of intonation contour requires detailed analysis not only on the acoustic but also on the linguistic level. The same utterance repeated several times by the same speaker may have various realization in terms of intonation (see, for example, Demenko 1999).

The examples in Fig. 1 and 2 show one kind of nuclear tune increasing length in terms of the number of syllables on one such series: 1) *znów*, (2) *znowu*, (3) *znowu on*, (4) *znowu ona*, 4) *znowu ten wariat* ("again" [short form], "again" [long form]; "it's him again"; "it's her again"; "it's that fool again"), always with the accent on the first syllable. The figures show these phrases as (a) spoken by the model and (b) as imitated in the experiment, (1) representing the low fall (ML) and (2) representing the high rise (LH). Physical comparison of intonation contours of the same phrase made by a given speaker shall indicate different character and range of pitch variations.

The issue of the imitation of certain intonation types and their application in speech recognition constitutes the subject of separate studies (see, for example, [Ashour & Gath 1999]).

Statistic assessment calls for representative and extensive linguistic material and to a large extent ignores complex relationships of pitch changes, the tempo of speech and the level of signal. Suprasegmental statistic parameters do not require detailed analysis of the melodic structure of speech and, consequently, they seem to be relatively easy to implement practically in automatic systems. The studies devoted to long-term suprasegmental analysis focus mainly on statistic features of the distribution of fundamental frequency (see, for example, Jassem & Dobrogowska, 1980).

The issues associated with the establishment of long-term variability in people's voices and the integration of segmental and suprasegmental features are essential for practical implementation but so far they have not been the subject of extensive research.

The purpose of this project is the preliminary assessment of the usefulness of both dynamic and statistic suprasegmental speech features in voice recognition.

III. EXPERIMENTAL PROCEDURE

3.1. Experimental material

The linguistic material included read story, read dialogue and spontaneous speech (samples of speech of the approx. length of 3 minutes). Fifty speakers (22 females and 18 males, aged 12-40) were recorded 3 times. Following preliminary perceptive analysis, three persons were excluded (due to numerous linguistic errors). Computer speech station Kay 5500 was used for basic acoustic analyses. Speech parameters such as fundamental frequency, intensity and jitter were established with using of software package Praat (Boersma & Weenink, 1996).

3.2. Dynamic analysis of suprasegmental features

The linguistic material adopted comprises 30 statements which differ with respect to syntax and semantics. Perceptive and acoustic analysis has identified nuclear tune ML in on average over 80% of this material, in other cases nuclear tune HL appeared. Fig. 3 presents exemplary realization of the HL tune in the word "odebrałas"(answered) from the sequence *mam nadzieje ze odebrałas telefon od Ali, (I hope that you answered Ala's call)* extracted for 4 persons. In the Fig. 3a one can notice the increase of all 3 parameters for the stressed vowel: F₀, duration and intensity (in comparison with the neighboring unstressed vowels *e* and *a*). In the fig. 3b stressed vowel is characterized by the increase

of the value of the F_0 parameter and duration, however, it has much lower intensity than the preceding unstressed vowel. Fig. 3c shows the example of the increase of the value F_0 and the intensity in case of the stressed vowel. Fig. 3d presents the increase of F_0 for the stressed vowel, while the intensity and duration have lower values in comparison to neighboring vowels. For each speaker, for selected sentences with the nuclear tone HL, the value of the F_0 parameter, duration and intensity on the stressed vowel, the vowel preceding the stressed vowel and the vowel following the stressed vowel were determined. In the case of all speakers, there was an increase on the stressed vowel of the fundamental frequency in comparison to neighboring vowels. In the case of 60% of persons, the stressed vowel was prolonged (on average by 20% in comparison to the preceding vowel). No observable changes of intensity were found.

Also statistical features of the whole sentence were subject to preliminary analysis. The linear trend (which reflects declination) was analyzed in its relation to the fundamental frequency in particular sentences. For the set of 30 sentences, 2 parameters of the linear function were determined for each speaker. Vectors of features were determined for the input of the MLP-type neural network. The vectors were defined by means of 4 values: the range of variations of F_0 parameter, minimum value of F_0 parameter (for each individual sentence) and the two parameters of the linear function. On average there were 72% of correct voice recognition.

3.3. Statistical analysis of continuous speech.

For each speaker 6 fragments of continuous speech (of the duration approximating 30 seconds) were considered. The number of observations in each sample depends on the ratio of the voiced segments to the voiceless segments of the speech wave. This number varied and was in most samples between 1100 -1500. Fig. 4 presents, for example, 4 samples for 2 female and 2 male voices. One can notice that irrespective of the selected text, the statistical features are relatively stable (as it is indicated by particular approximations of the four lognormal distributions).

The following parameters were subject to statistical analysis: mean, minimal, maximal, initial value of fundamental frequency, standard deviation, skewness and kurtosis of the distribution. Also the micro variations of fundamental frequency-jitter were subject to analysis which, as it has been proved (see, for example, Schoentgen, 1999), includes also person-specific information. The relations between voiced and unvoiced phonetic-acoustical speech segments were also analyzed since some speakers have difficulties with production of certain voiced consonants.

3.4. Results

Preliminary descriptive statistics and the Analysis of Variance of the variations of selected parameters have shown that the minimal value of the F_0 parameter (varying in the range 4.5-5.6) and the average value (varying in the range 4-4.8) have the greatest share in individual voice characteristics. These parameters of a given voice are relatively stable from the statistical point of view. Also pitch variability range (0.7-2.4) and standard deviation (0.08-2.0) proved to be important. The ratio of the number of voiced to voiceless samples (D/B - Voiced/Voiceless) equaled on average 47% (it varied depending on the speaker in the range 39-66%). Some speakers have the tendency to devoice consonants in the positions in which voiced consonants appear in normative speech. Jitter (varying from 0.6% - 2%) also proved to be important. Wide statistical spread (standard deviation up to 0.3) denoted initial values of the courses and maximum fundamental frequencies, hence these parameters seem of little application for the assessment of individual voice features.

Table 1 presents the results of discriminant analysis, the evaluation of the usefulness of analyzed features in the recognition of 47 voices under examination.

Discriminant analysis recognition results (on average 76% correct recognitions) seem to indicate that there is a possibility of using also long-term features in voice recognition. However, the analysis should be extended and the technique of neural networks should be applied which do not require assumptions as to the normality of distribution.

Features	Main Wilks' Lambda	Wilks' Lambda	F statistic	Level of significance
D/B	0,000016	0,530488	2,57820	0,000000
Jitter	0,000014	0,493400	2,65688	0,000000
Average	0,000075	0,115100	22,39578	0,000000
Range	0,000050	0,171178	14,10459	0,000000
Min	0,000092	0,093500	21,24109	0,000000
Standard deviation	0,000037	0,233283	9,574100	0,000000

Table 1. Results of the discriminant analysis.

IV. SUMMARY.

The results of the analyses conducted within a sentence and in continuous speech indicate significant individual differentiation both in respect of statistic parameters and of dynamic suprasegmentals. However, the assessment of their practical usefulness in automatic voice recognition systems requires extending experimental material and solving the problem of the integration of selected suprasegmental features with segmental features. This is the objective of the currently conducted project /ARG/ *Automatic Voice Recognition Systems for the Blind* which is carried out on the basis of an extensive data base of recordings of continuous speech as well as isolated utterances.

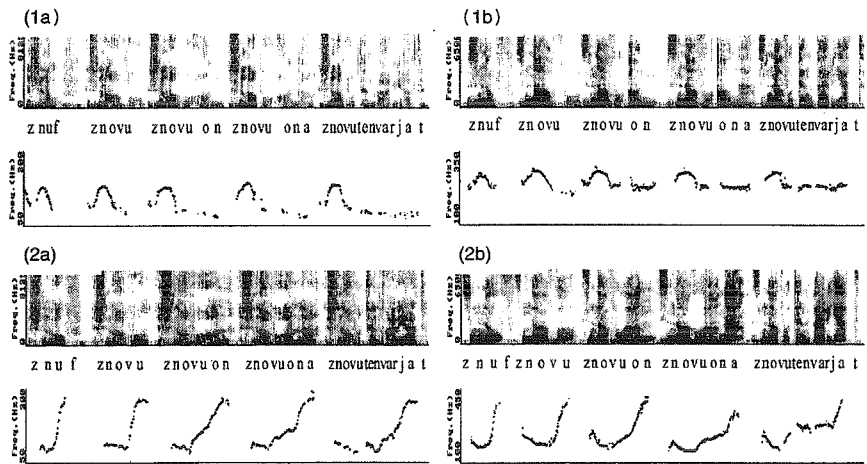


Fig.1a. ML ("again" [short form], "again" [long form]; "it's him again"; "it's her again"; "it's that fool again") Fig.1.b. ML imitation
 Fig.2a. LH ("again" [short form], "again" [long form]; "it's him again"; "it's her again"; "it's that fool again") Fig.2b. LH imitation

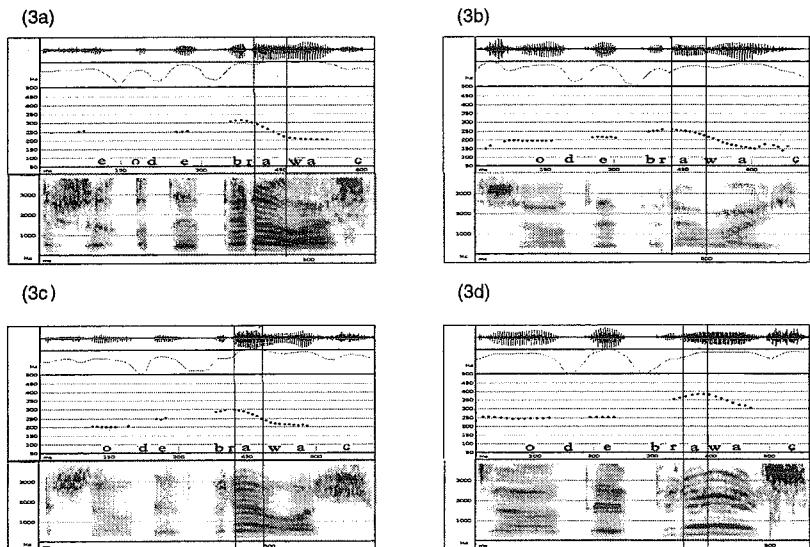


Fig.3. Realization of the nuclear tune HL in the word *odebrala* (answered). The stressed vowel was indicated by means of cursors. Fig. 3a. Speaker JK Fig.3b. Speaker RL
 Fig. 3c. Speaker TS Fig.3d. Speaker AB.

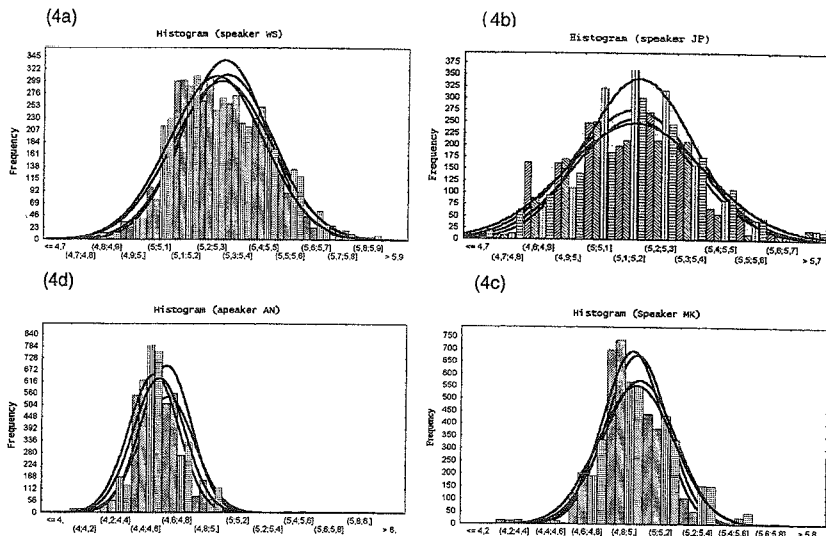


Fig.4. Distributions of Fo parameter for 4 different 30-second fragments of a text. Solid line indicates lognormal approximation of the distribution.

REFERENCES

- Ashour, G., Gath, I. (1999) "Characterization of Speech during Imitation", Proceedings of Eurospeech'99, v.3, 1187-1190.
- Atal B. S. (1972) "Automatic speaker recognition based on pitch contours", J.Acoust.Soc.Am. 52, 1687-1697.
- Barlow, M., Wagner M. (1998) "Measuring the Dynamic Encoding of Speaker Identity and Dialect in Prosodic Parameters", Proceedings of ICSLP'98, vol.2, 81-85.
- Boersma, P., Weenink D. (1996) "Praat, a system for doing phonetics by computer", Report 132 of the Institute of Phonetic Sciences Amsterdam.
- Demenko, G. (1999) "Analysis of suprasegmentals for speech technology", wyd.UAM, Poznań.
- Jassem W., Dobrogowska K. (1980) "Speaker-Independent Intonation Curves, in The melody of Language", University Park Press, Baltimore, 135-149
- Jassem W., Demenko G. (1986) "On Extracting Linguistic Information from F0 traces", in Intonation in Discourse, Croom Helm, London, 1-18.
- Mathew, M., Yegnanarayana B., Sundar R. (1999) "A neural Network-Based Text-Dependent Speaker Verification System Using Suprasegmental Features", Eurospeech'99, vol.2, 995-998.
- Schoentgen, J.(1999) "A random-walk model of jitter", Proceedings of the ICPHS99, San Francisco, 2441-2444.

ACKNOWLEDGEMENT

The study has been performed within the 8T11 E022 17 research project financed by KBN.