

Spectral Characteristics of Vocal Tract for Speaker Recognition

Milan Sigmund

Institute of Radio Electronics, Brno University of Technology, Purkynova 118, CZ-61200 Brno, Czech Republic

Summary

The basic idea of the presented approach is to evaluate a spectral characteristics corresponding to the anatomy of the speaker's vocal tract independently of the actually pronounced phoneme. The procedure for determining the speaker-specific average spectrum is based on the LPC approach. Experimental results have shown an evolution in a long-time spectrum with respect to the duration of text independent utterance. This was compared with a representative set of chosen isolated words using a long-time spectrum variability between and within speakers. The evaluated method is suitable for both text-dependent and text-independent speech signal processing.

Key words:

Speaker recognition, individual spectral characteristics.

1. Introduction

Acoustical communication is one of the fundamental prerequisites for the existence of human society. Textual language has become extremely important in modern life, but speech has dimensions of richness that text cannot approximate. About 25 % of the information contained in a speech signal refer to speakers. These phonetically irrelevant speaker properties make speech recognition less effective, but they can be used for speaker recognition. This is a fascinating area of speech research. Using only the speech fairly accurate estimations can be made as to whether the speaker is male or female, adult or child. Belligerence, anger, fear, sadness or joy may all be detectable in the speech signal. At present, the interest in this area of research is increasing as the number of potential applications increases and vocal emotions have also tended to be studied in an isolated way [1].

Long-time sample statistics of various spectral features are often used as text-independent features such as the mean and variance of spectral features over a series of utterances. However, long-time spectral averages are extreme condensations of the spectral characteristics of a speaker's utterances. In a long-time average spectrum of a speech signal the linguistic information (coded as frequency variation with time) is lost while the speaker specific information is retained. In this study, a speaker analysis approach based on linear predictive coding (LPC) is presented. The basic idea of the approach is to evaluate an average long-time spectrum corresponding to the anatomy of the speaker's vocal tract independent of the

actually pronounced phoneme. This approach is independent of language and very simple to implement.

2. Used Method

The procedure for determining the spectral characteristics of vocal tract is based on the LPC approach. First, we compute the short-time autocorrelation coefficients $R_j(k)$, $k=0, \dots, K$ for the j -th frame (20 msec) of speech signal $s(n)$

$$R_j(k) = \sum_{n=1}^{N-k} s(n)s(n+k) \quad (1)$$

where N is the number of samples of each frame, and then we compute the average autocorrelation coefficients

$$\bar{R}(k) = \frac{1}{J} \sum_{j=1}^J R_j(k) \quad (2)$$

corresponding to the whole utterance formed by J frames. Thus, from the average autocorrelation coefficients, we get the average predictor coefficients \bar{a}_m e.g., via the Durbin algorithm [2] and finally the normalised average LPC-based spectrum using

$$S(f) = \left| \frac{1}{1 - \sum_m \bar{a}_m z^{-m}} \right|_{z = \exp(j2\pi f / f_s)}^2 \quad (3)$$

for $m = 1, \dots, M$

where f_s is the sampling frequency and M is order of the LPC model equal to the highest autocorrelation value K . More details how to compute the LPC coefficients and corresponding spectra on short frame of speech signal can be found in [2].

The speech data used in the experiment described below were recorded with an electret microphone, held 15-20 cm from the lips. The speech signal was sampled at 22 kHz using a 16-bit A/D converter under laboratory conditions over a period of five months. A group of 26 speakers (19 male, 7 female) aged 20 to 25 years took part in the research, the speaker's nationalities were Czech and Hungarian.

3. Experiments and Results

3.1 Speech Duration

An important factor for the accuracy of vocal tract spectrum estimation is the needed speech duration. Duration refers to how much of the training/test data must be used to eliminate the text-dependent effect on the variation of the average spectrum. As an example, we present spectrum accuracy as a function of speech duration in Figure 1 for LPC order 6.

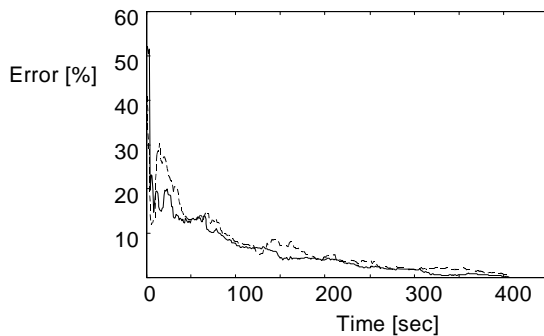


Fig. 1 Long-time spectrum accuracy as a function of the speech duration.

The solid and dotted curves correspond to the Czech and the Hungarian text spoken by the same speaker (native Hungarian living in the Czech Republic). Both curves differ in details but tend to the same contour.

3.2 Long-Time Spectrum Variability

For practical applications, it is desirable that a speaker recognition method should require only a small amount of training/test data. One way of using shorter speech signal duration consists in choosing a suitable set of testing data. For this purpose, an informal experiment was performed with a small specially isolated-word vocabulary of 10 most representative words of the Czech language chosen and recommended by the Institute of the Czech Language of the Czech Academy of Sciences. The Repe-Set reflects the most significant phonetical features of Czech:

- they are phonetically balanced words,
 - the complete phoneme repertory is covered,
 - the words are of type CVCVC (C=consonant, V=vocal), which is the common structure of Czech words.
- The text-independent data, on the other hand, was read from newspaper texts.

Figure 2 illustrates an estimation of the long-time spectrum obtained by means of the representative set of words (solid line), which represents about 10 seconds of

speech. The other curves correspond to free texts of various duration: dotted - 10 sec, dot-dashed - 50 sec, dashed - 100 sec.

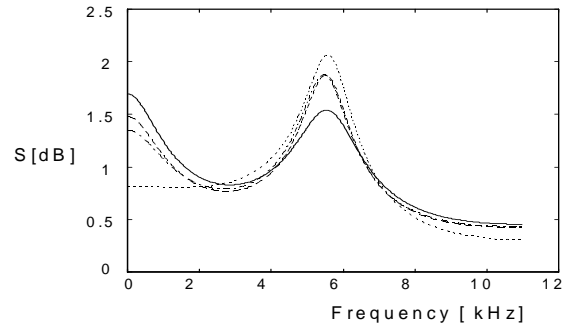


Fig. 2 Comparison of long-time spectra obtained from free texts of various duration and from a selected text spoken by the same speaker.

A comparison between intra- and inter-speaker variability in long-time spectrum is shown in Figures 3 and 4. Figure 3 illustrates two vocal tract spectra of the same speaker corresponding to two different texts. The difference between both curves is 12%.

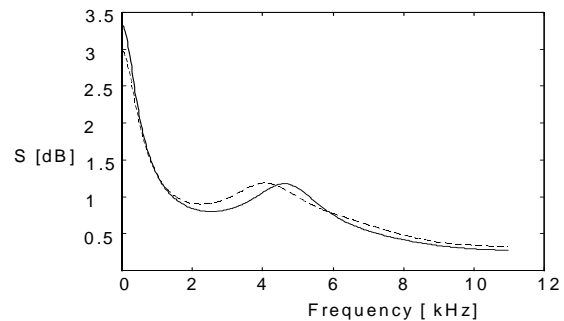


Fig. 3 Long-time spectrum difference of one and the same speaker (LPC order 6, speech duration 100 sec).

Vocal tract spectra obtained from two different speakers saying the same text is shown in Figure 4. The difference between both curves increased to 22% in this case. The average intra-speaker difference over all speakers was 12.6%, while the average inter-speaker difference (gender-specific) reached 23.4%. In accordance with the inter-gender differences, the estimated difference between the two groups of speakers (male and female) was more apparent (29.6%) than within the groups [3].

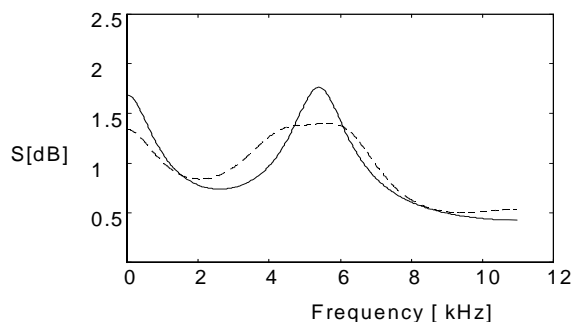


Fig. 4 Long-time spectrum variability between speakers (speech duration 100 sec).

The effect of changes in speech due to the emotional state of speaker on long-time spectrum can be observed in Figure 5.

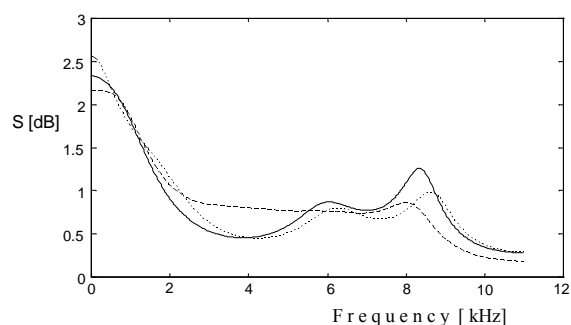


Fig. 5 Long-time spectrum variability within speaker for normal and emotional speech (speech duration 114 sec).

The dashed line gives the spectrum of emotional speech spoken under stress, the solid line gives the spectrum obtained from the same text read in normal state of speaker and the dotted line also gives the spectrum from the same text read by a tired speaker. Thus in all three cases the identical speech was spoken by one speaker in various states of mind. The changes in psychological state (stress) affect the spectrum more than the changes in physical state (fatigue). For our studies conducted within the research of speech processing under stress we used our own database [4] consisting of speech data collected during oral final examinations at our Institute of Radio Electronics.

4. Conclusions

In this paper, a new estimation of spectral characteristics of vocal tract by average long-time spectrum and some contributions to speaker recognition are presented. In summary, the following conclusions can be drawn from the experiments and data mentioned in this work. It is possible to use long-time spectra models across languages for normally spoken speech. To estimate relevant long-time spectra with respect to their computational simplicity, speech of about 100 seconds in duration seem to be sufficient. Long-time spectra can yield high speaker recognition accuracy for normal speech but not for speech spoken under stress and for disguised (impersonated) speech. In addition, representative small subset of 10 most representative words of the Czech language was successfully tested.

Acknowledgment

This work was supported by the Research Programme of Brno University of Technology No. MSM 0021630513 Advanced Electronic Communication Systems and Technologies.

References

- [1] Baken, R. J. and Orlikoff, R. F.: Clinical Measurement of Speech and Voice, Singular Publishing Group, San Diego, 2000.
- [2] Rabiner, L. and Juang, B.: Fundamentals of Speech Recognition, Englewood Cliffs, New Jersey, 1993.
- [3] Sigmund, M. and Menšík, R.: Estimation of Vocal Tract Long-Time Spectrum, in Proc. Elektronische Sprachsignalverarbeitung, Dresden, Germany, 1998, pp. 69-71.
- [4] Sigmund, M. and Severňák, O.: Eine neue Sprachdatenbank mit der Sprache unter Stress, in Proc. Elektronische Sprachsignalverarbeitung, Bonn, Germany, 2001, pp.323-328. (in German)



Milan Sigmund received a masters degree in 1984 in biomedical engineering and a doctoral degree in 1990 in speech signal processing, both from the Brno University of Technology (Czech Republic). Currently, he is in the Faculty of Electrical Engineering and Communication at Brno University of Technology. In the years from 2001 to 2003, he was in the

Department of Computer Science at the University of Applied Sciences Wiesbaden (Germany). His main research interests include speech signal processing with a special focus on automatic speaker recognition. He is a member of ISCA and EAEEIE.