Journal of ELECTRICAL ENGINEERING, VOL. 57, NO. 5, 2006, 301–303

COMMUNICATIONS

SPECTRAL MODELS OF INDIVIDUAL SPEAKERS

Milan Sigmund *

The basic idea of the presented speaker analysis approach is to evaluate the spectral model corresponding to the anatomy of the vocal tract of the speaker independent of the actually pronounced phoneme. The procedure for determining the speaker-specific average spectrum is based on the linear predictive coding (LPC) approach. Experimental results show the evolution of a long-time spectrum with respect to the duration of the text independent utterance compared with a specially isolated word vocabulary chosen for the Czech language, a long-time spectrum variability between and within speakers, the effect of the emotional state of the speaker on the long-time spectrum and speaker normalisation by the long-time spectrum. The method is suitable for both text-dependent and text-independent tasks.

K e y w o r d s: speaker recognition, individual long-time spectrum

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 information contained in speech signal refer to speaker. These phonetically irrelevant speaker factors make the speech recognition less effective, but they can be used for speaker recognition. This is a fascinating area of speech research. From speech alone, fairly accurate guesses can be made as to whether the speaker is male or female, adult or child. Belligerence, anger, fear, sadness or elation may all be detectable in the speech signal. At present, interest in this area of research is increasing as the number of potential applications grows and vocal emotions have also tended to be studied in isolated nature [1].

Used as text-independent features are often long-time sample statistics of various spectral 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 vocal tract of the speaker independent of the actually pronounced phoneme. It is independent of language, and very simple to implement.

2 USED METHOD

The procedure for determining the speaker-specific average spectrum is based on the LPC approach. First, we compute the short-time autocorrelation coefficients $R_j(k),\ k = 0, 1, ..., K$ for the j-th frame of speech signal $s(n)$.

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

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

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

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

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

for $m = 1, \ldots, M$, 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

* Institute of Radio Electronics, Brno University of Technology, Purkynova 118, CZ-61200 Brno, Czech Republic, E-mail: sigmund@feec.vutbr.cz

ISSN 1335-3632

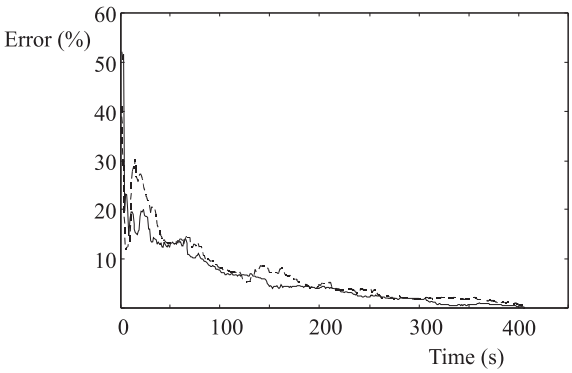

Fig. 1. Long-time spectrum accuracy as a function of the speech duration

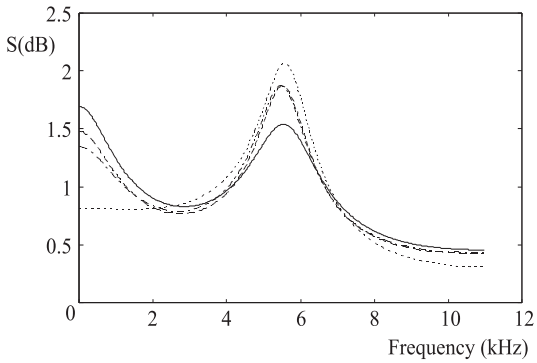

Fig. 2. Comparison of long-time spectra obtained from free texts of various duration and from a selected text spoken by the same speaker

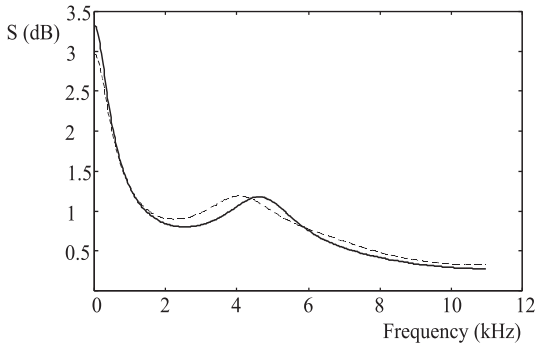

Fig. 3. Long-time spectrum difference of one and the same speaker (LPC order 6, speech duration 100 s)

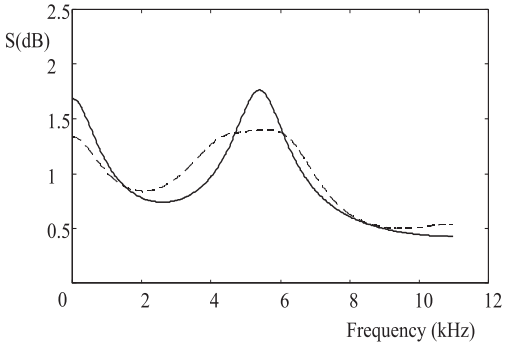

Fig. 4. Long-time spectrum variability between speakers (speech duration 100 s)

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.

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 Repre-Set (*duben, garáž, hořák, major, podíl, pochod, součet, tisíc, výfuk, žízeň*) 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)

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 s, dot-dashed - 50 s, dashed - 100 s.

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%.

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].

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 psychological state (stress) affects the spectrum more than the physical state (fatigue). For our studies conducted within the research of speech

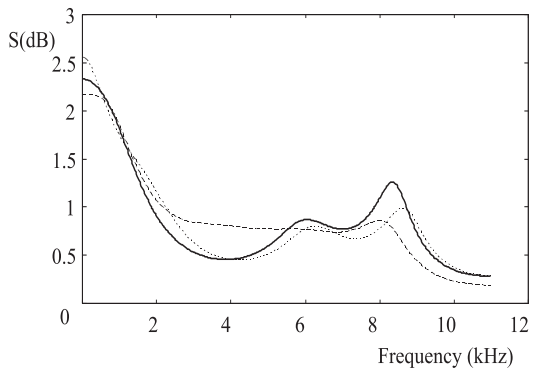

Fig. 5. Long-time spectrum variability within speaker for normal and emotional speech (speech duration 114 s)

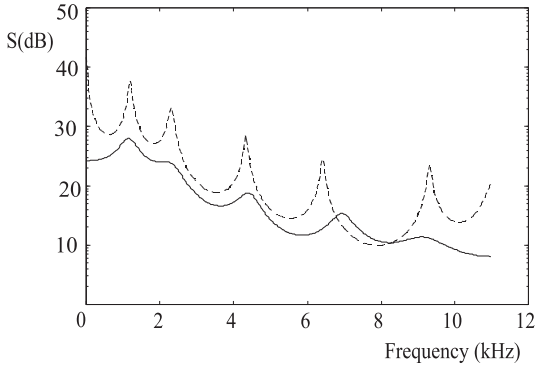

Fig. 6. Effect of speaker normalisation on the spectral function for phonemes "a". Solid line shows the spectrum before normalisation, dotted line after normalisation.

processing under stress we used our own database [4] consisting of data collected during oral final examinations at our Institute of Radio Electronics.

3.3 Speaker Normalisation by Long-Time Spectrum

An important aspect of the described long-time spectrum is that it also offers a potential tool for speaker normalisation [5] applicable to speaker-independent continuous speech recognition. Figure 6 illustrates the effects of the normalisation by long-time spectrum using for the spectrum of vowel "a" cut out from continuous speech. The formant peaks of normalised speech are weighted more heavily and thus represented more accurately.

4 CONCLUSIONS

In this paper, a new estimation of speaker characteristics 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. Long-time spectra used for speaker normalisation can bring better formant localisation and increased performance of word recognition systems. In addition, representative small subset of 10 most representative words of the Czech language was successfully tested.

The increase in application opportunities has resulted in increased interest in voice recognition research. Speaker recognition is nowadays regarded by market projections as one of the more promising technologies of the future. We look forward to seeing the promise of speech and speaker recognition become a reality.

Acknowledgments

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.—ORLIKOFF, R. F.: Clinical Measurement of Speech and Voice, Singular Publishing Group, San Diego, 2000.

[2] RABINER, L.—JUANG, B.: Fundamentals of Speech Recognition, Englewood Cliffs, New Jersey, 1993.

[3] SIGMUND, M.—MENŠÍK, R.: Estimation of Vocal Tract Long-Time Spectrum, in Proc. Elektronische Sprachsignalverarbeitung, Dresden, Germany, 1998, 69-71.

[4] SIGMUND, M.—SEVERŇÁK, O.: Eine neue Sprachdatenbank mit der Sprache unter Stress, in Proc. Elektronische Sprachsignalverarbeitung, Bonn, Germany, 2001, 323-328. (in German)

[5] SIGMUND, M.: Speaker Normalisation by Long-Time Spectrum, in: Proc. Radioelektronika'96, Brno, Czech Republic, 1996, 144-147.

Received 30 September 2005

Milan Sigmund (Assoc Prof, Ing, CSc) was born in Ivančice (South Moravia). He received a masters degree in 1984 in biomedical engineering and a doctoral degree in 1990 in speech signal processing, both from the Technical University of Brno (Czech Republic). Currently, he is on 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. His main research interests include speech signal processing with a special focus on automatic speaker recognition. He is a member of the International Speech Communication Association (ISCA) and a member of the European Association for Education in Electrical and Information Engineering (EAEEIE).