Spectral Analysis of Speech under Stress

Milan Sigmund,

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

Summary
This paper deals with speech signal as significant indicator of psychological stress when the speaker is involved in a stressful activity. The investigation of speaker's stress is based on specific changes in short-time spectrum of vowel phonemes. For each selected signal segment, the spectrum is computed by means of two different methods: Fourier transformation and chirp transformation. Comparative results between both spectra serve for speaker's stress detection. In case of speech under stress, the obtained spectra differ towards the higher frequencies due to enhanced pitch modulation observed in the envelope of the chirp spectrum. For use in our experiments, a new database of speech under stress consisting of data collected during oral examinations at our university was created.

Key words:
Emotional speech signal, spectral analysis.

1. Introduction

Emotions have long been recognized to be an important aspect of human beings. More recently, psychologists have begun to explore the role of emotions as a positive component in human cognition and intelligence. Spoken language comes from our inside. Factors such as mood, emotion, physical characteristics and further pragmatic information are contained in speech signals. Many of these characteristics are also audible. An emotional speech with high content differs in some parameters from a neutral speech [1]. In recent years, the interest for automatically detection and interpretation of emotions in speech has grown and vocal emotions have also tended to be studied in isolation. About 25% of information contained in a clean speech signal refer to the speaker. These linguistically irrelevant speaker characteristics make speech recognition less effective but can be used for speaker recognition (ca. 15% of information) and analysis of the speaker's emotional and health state (ca. 10% of information).

With increasing demand for speech technology systems, there is an increasing need for processing of emotion and other pragmatic effects (simulation in synthetic speech, elimination in robust speech recognition). In some cases, it is very important to detect the emotional state of a person (e.g., stress, fatigue or use of alcohol) from his/her voice.

Manuscript received April 5, 2007
Manuscript revised April 25, 2007

2. Speech under Stress

Stress is a psycho-physiological state characterized by subjective strain, dysfunctional physiological activity and deterioration of performance. The accepted term for speech signal carrying information on the speaker's physiological stress is "stressed speech".

2.1 Speech Production under Stress

Speech production begins with abstract mental processes: the desire to communicate and the idea which is to be communicated. Suitable linguistic units have to be chosen from memory and formed into a sentence, subject to grammatical constraints. From the abstract sequence of words, a corresponding sequence of articulatory targets must be generated. Then appropriate motor programs for the targets must be activated, with modifications to take account of context and paralinguistic information. This results in patterns of nerve impulses being transmitted to the muscles which control the respiratory system and vocal tract. The final stages are purely physical: the generation of acoustic energy, the shaping of its spectral and temporal characteristics, and its radiation from the mouth.

2.2 Available Speech Databases

There are not many corpora designed to allow the study of speech under stress. A typical corpus of extremely stressed speech from a real case is extracted from the cockpit voice recorder of a crashed aircraft. On the other hand, such extreme situations occur seldom in everyday life. The mostly mentioned corpus in the literature is the SUSAS (Speech Under Simulated and Actual Stress) database of stressed American English described in [2]. An overview of emotional speech data collections (including stressed speech) is presented in [3]. There are mentioned the significant European languages, Japanese, and Chinese. For the research at our university, the most suitable situation with realistic stress took place during the final state examinations held in oral form in front of a board of examiners. The test speakers in our experiment were 31

male students from own database called "ExamStress". A complete description of this database can be found in [4].

3. Spectral Analysis

To get the quantitative changes of speech parameters, we applied some simple features that had not been specifically designed for the detection of stressed speech. First of all, the speech signals were investigated in the frequency domain. The spectral structure of speech is usually observed by means of Fourier analysis. However, in case of emotional intonation, which gives rise to continuous and fast variation of the pitch, the harmonic structure is very often "hidden" [5]. In order to enhance the spectral structure, the so-called chirp transformation was tested.

The discrete short-time chirp transformation of signal $s(n)$ is defined as

$$S(k,\alpha)=\sum_{n=0}^{N-1}s(n)\sqrt{\Phi_\alpha(n)}\,\exp(-j2\pi\frac{k}{N}\Phi_\alpha(n)\,, \quad (1)$$

where k is the frequency index, N is the length of segment (number of segment samples) and $\Phi_\alpha(n)$ is the following mapping, bijective in the interval of interest <0, N>

$$\Phi_\alpha(n)=\left[1+\frac{1}{2}\alpha(n)\right]n, \quad (2)$$

where parameter α is the so-called chirp rate. The bijectivity of $\Phi_\alpha(n)$ results in the following limits for the chirp rate α

$$-\frac{2}{N}<\alpha<\frac{2}{N}\,. \quad (3)$$

The chirp transformation is a generalization of the Fourier transformation, which corresponds to α=0. Detail description of the chirp transformation can be found in [5].

4. Results

In our experiments, we used phonetically rich sentences from the ExamStress corpus for our analysis of stressed speech. These sentences were automatically segmented into phoneme-like units. As expected, some sounds represent the speaker individuality better than others. Our previous experiments show that vowels and nasals are the most suitable phonemes for speaker analysis [6]. They are also relatively easy to identify in speech signals. In terms of speaker-recognition power, the following ranking of phoneme classes results (in decreasing order):

vowels, nasals > liquids > fricatives, plosives

The most effective individual phonemes for speaker investigation seem to be "e" and "a"; in string with some nasals or liquids they can reliably characterize the speaker. Using the FonLabel program [7] developed for speech segmentation, we can automatically select vowel phonemes and for each vowel derive its acoustic features.

In this section, some spectral features are briefly presented. These features were measured in both neutral and stressed speech, and obtained values were then compared. In all experiments, the speech signal was digitized at 8 kHz, sampling rate with 8 bit resolution and subsequently multiplied by the Hamming window function. Thus, the analyzed frequency range is 0-4000 Hz which corresponds to spectrographic methods usually used in the forensic speaker recognition. The analyzed short-time speech frames spanned 30 ms. In the search for optimal signal representation in the frequency domain various spectral transformation were applied to the speech data. Finally, the spectra were computed using the Fourier transformation and chirp transformation. In case of neutral speech, the vowel spectra obtained from the same speech frame have very similar envelope. This situation is illustrated in Fig. 1.

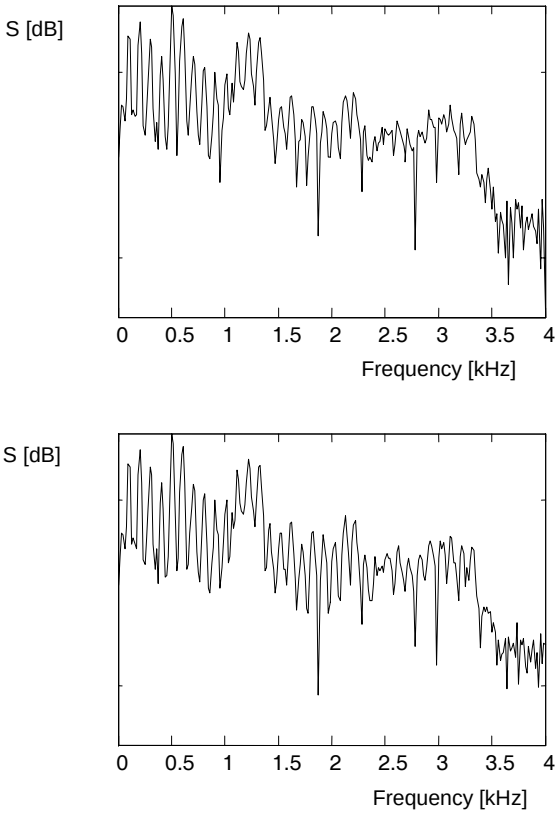

Fig. 1 Examples of spectrum of the phoneme "e" obtained from neutral speech by means of Fourier transformation (upper graph) and chirp transformation (lower graph).

On the contrary, both spectra differ in case of stressed speech. The chirp transformation delivers a very fine spectral representation of vowels with enhanced pitch modulation. A visual comparative between both spectra shows that the main differences lie in the frequency band 2-3 kHz. These effects are obvious for almost any speaker. An example of such typical spectral envelopes can be seen in Fig. 2. It should be noted that all spectra shown in Fig. 1 and Fig. 2 were computed from phoneme "e" extracted from identical word spoken by the same speaker.

In order to compare the Fourier and chirp spectrum automatically, it is inevitable to find a useful mathematical description. An effective criterion seems to be a distance measure derived from a correlation based similarity measure

$$d = 1 - \frac{< \mathbf{x}, \mathbf{y} >}{\|\mathbf{x}\| \, \|\mathbf{y}\|}, \qquad (4)$$

where **x** and **y** represent the spectral vectors.

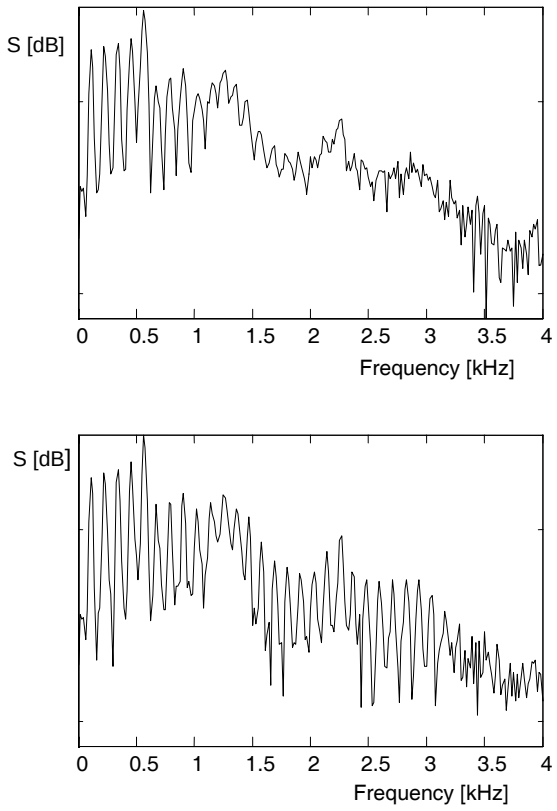

Fig. 2 Examples of spectrum of the phoneme "e" obtained from speech under stress by means of Fourier transformation (upper graph) and chirp transformation (lower graph).

5. Conclusion

Changes in spectrum of speech signal have shown to be a reliable indicator of the internal emotional state of a person. The speaker's stress can be detected from short segments of vowels comparing their Fourier and chirp spectra. Our long-term goal is to automatically detect and quantify the actual stress influencing a person, on the basis of acoustic and prosodic information extractable from utterances.

Acknowledgments

This work was supported by the Research Plan of Brno University of Technology No. MSM 0021630513 "Advanced Electronic Communication Systems and Technologies (ELCOM)".

References

[1] T. Johnstone and K. Scherer, "The effects of emotions on voice quality," *Proceedings of 14th International Congress of Phonetic Science*. San Francisco, pp. 2029-2032, 1999.

[2] J. H. Hansen and S. E. Ghazale, "Getting started with SUSAS," *Proceedings of Eurospeech'97*. Rhodes, pp. 1743-1746, 1997.

[3] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication,* vol. 48, No. 9, pp. 1162-1181, 2006.

[4] M. Sigmund, "Introducing the database ExamStress for speech under stress," *Proceedings of 7th IEEE Nordic Signal Processing Symposium (NORSIG 2006)*. Reykjavik, pp. 290-293, 2006.

[5] M. Képesi and L. Weruaga, "Adaptive chirp-based time-frequency analysis of speech signals," *Speech Communication,* vol. 48, No. 5, pp. 474-492, 2006.

[6] M. Sigmund, *Voice Recognition by Computer*. Tectum Verlag, Marburg, 2003.

[7] M. Sigmund and P. Matějka, "An environment for automatic speech signal labeling," *Proceedings of 28th IASTED International Conference on Applied Informatics*. Innsbruck, pp. 298-301, 2002.

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 stayed 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.