SPEAKER STRESS DETECTION BY ANALYSIS OF GLOTTAL EXCITATION

M. Bostik, M. Sigmund
Department of Radio Electronics, Brno University of Technology, Brno, Czech Republic

Abstract: In this contribution the recognition of stress and emotional state is analysed by speech signal analysis using Liljencrant-Fant’s model. It is based on the knowledge that some parameters of glottal pulses, obtained by this model, are changed owing to stress, hence they are suitable for the detection of speaker’s stressed (“abnormal”) state. Two procedures for the analysis of these parameters are described in detail. The first of them is an analysis of parameters of randomly chosen speech parts (of phonetically constant length) that makes fewer demands on segment selection, the second is an analysis of speech parts going one by one in time. The methods were applied to sound recordings made at “stressed” oral examinations at a university. The results obtained show the applicability of these parameters and methods especially for speech analysis when we have at our disposal a signal recorded in the “normal” (steady) state of speaker.

Keywords: stress, glottal excitation

I. INTRODUCTION

The usually used methods for identifying stress and other emotional states [1] usually start from the time distribution of single phonetic parts of words or sentences. Speech influenced by psychical stress can be identified e.g. by different time lengths of the concrete phonemes or by different time lengths of speech pauses between words [2], [3]. Statistical evaluation is also often used to examine e.g. the distribution function of the first two formants or the distribution of time samples. Also used are classifiers based on the pitch period detection and its variation in time. All procedures described above have one common factor, namely that a long time record has to be processed (for statistical methods it is necessary). In the present contribution the method of recognizing stress and some other emotional states, based on the analysis of one or a few period of speech signal is discussed. The low time requirements (from the viewpoint of the length of speech signal not computation) of the method are paid for by the need to own a sound record of the speaker at “normal” state and if it is possible of the some phonetic content. The description of the analysis of the speech signal using Liljencrant-Fant’s (LF) model can be found in [4]. This model estimates parameters of glottal pulses ($E_c$, $\omega_c$, $\alpha$ and $\varepsilon$ in Fig. 5) and can also be used for speech signal synthesis and the parameters of this model it is possible to imitate the voice of a specific person. Some parameters of glottal pulses, obtained by the LF model, are especially suitable for “abnormal” speaker state identification [5]. In the following sections two procedures for processing the obtained LF parameters are described and their results are compared in the conclusion. The first procedure is an analysis of the parameters computed from randomly chosen parts (of phonetically constant length) that makes fewer demands on the selection of segments, and the second procedure is an analysis of the parameters obtained from speech parts going one by one in time. The methods were applied to sound recordings made at a diploma work defence, under the influence of speakers’ examination stress.

II. SPEECH DATA

It is really difficult to obtain realistic voice samples of speakers in various stressed states, recorded in real situations. There are not many corpora designed to allow the study of speech under stress. A typical corpus of stressed speech from a real case is extracted from the cockpit voice recorder of a crashed aircraft. The only publicly available corpus is the SUSAS database of stressed American English. Two of our own databases [10] were created for use in our experiments; a database of stressed speech and a database of alcoholic speech.

However, for our studies conducted within the research of speech processing in noise and stress we used our own database, namely the SZZ database, consisting of data collected during oral final examinations at our Institute of Radio Electronics. The recorded utterances were manually examined (including both examination of the waveform and parameter contour, and listening) and then endpoints of words were determined. In this way, a number of pauses and irrelevant extraneous voices were eliminated. This material contains stressful phases (improvisations relating to unknown technical problems) and other phases with lower stress (during discussions relating to known technical problems). The hardware and software were hosted by a PC hooked up to the local net for automatic backup of the recorded speech files. The recording platform is set up to store the speech signals “live” in 16-bit coded samples at a sampling rate of 22 kHz. Thus, the acoustic quality of the records is determined by the speaking style of the students and the background noise in the room. For the experiment, only voiced speech segments were used because of our previous experience.
III. METHODS USED

Fig. 1 shows the block diagram of a system for obtaining the LF parameters from continuous speech.

![Fig. 1 Block diagram for the LF parameters estimation.](image)

The function of the single blocks in Fig. 1 is described in detail in the following.

- **Segment selection**

Before we can start to analyse speech segments it is necessary to choose, by some suitable method, speech signal parts that are suitable for analysis. It is, for example, unsuitable to analyse unvoiced parts if our aim is to obtain and to describe glottal pulses of the vocal apparatus. The next criterion for the selection can be e.g. the difficulty of selection and selection effectivity (effectivity is to be understood in this case as the ratio of the sum of time lengths of the chosen segments from a concrete set of speech data and the whole time length of the set). If we choose a concrete phoneme from the speech data, the selection effectivity is small and the time length of speech data increases (if we want to preserve the level of statistical reliability).

The main aim of this work is to find a suitable procedure for segment selection for “abnormal” speaker state identification. Two methods were used and tested. The first method assumes that the LF parameters of the glottal pulses are rather constant and do not change much during the speech due to coarticulation. Then it is possible to choose voiced segments randomly, independently of the position in the utterance, see Fig. 2.

![Fig. 2 Random segment selection, up – continuous speech, down – chosen segments situated one by one in time.](image)

The second method assumes that the LF parameters of the glottal pulses are changed during the speech. Then it is necessary to choose voiced segments one by one in time, in dependence on the position in the speech, see Fig. 3.

![Fig. 3 Segment selection one by one in time, up – continuous speech, down – chosen segments.](image)

A further limitation is that we have to have phonetically identical utterances of “normal” and “abnormal” speech.

- **Estimation of glottal pulse waveform**

For the glottal pulse estimation, several methods exist [5]. The well known and effective method is the transfer function estimation of vocal tract with subsequent inverse filtering. This algorithm is one of the basic methods of speech signal processing, further information can be found e.g. in [5], [6], another similar algorithm is presented in [7]. For illustration, Fig. 4 shows a primary speech signal and its excitation signal obtained by inverse filtering.

![Fig. 4 Time waveform, up – speech signal (phoneme “a”), down – speech signal after inverse filtering.](image)

- **Approximation using the LF model**

Now, it remains to mention the computation and properties of LF parameters. Glottal pulse approximation using the LF model uses, as the approximation curve, the exponential function combined with harmonic function. That can be seen in Eq. (1) and Eq. (2). Vectors $v_{g1}(n)$ and $v_{g2}(n)$ are two parts of the approximation curve and together they form approximation function $v_g$, see Fig. 5.
\[ v_{el}(t) = -E_e \frac{\sin[\omega_e (t - T_{op})]}{\sin[\omega_e (T_e - T_{op})]} e^{\alpha (t - T_e)} \]  
for \( T_{op} \leq t \leq T_e \) \hspace{1cm} (1)

\[ v_{eg}(t) = \frac{-E_e}{\epsilon T_a} e^{\epsilon (T_e - t)} - e^{\epsilon (T_e - T_a)} \]  
for \( T_e < t < T_a \) \hspace{1cm} (2)

Variables \( T_{op}, T_e, T_c \) and time interval \( T_a \) are important parameters and their meaning can be clear from Fig. 5. Approximation is limited to the time interval \( T_{op} \leq t \leq T_e \). The remaining variables \( E_e, \omega_g, \alpha \) and \( \epsilon \) are the LF parameters sought. It is possible to obtain them by some of the iterative methods. The parameters are determined by criteria of the minimal average quadratic deviation of the approximating and the approximated function. All procedures described here were implemented using mathematical software Matlab on the modified PC with a professional sound card.

**Fig. 5** Time waveform of the approximation function and the meaning of individual parameters.

### IV. DATA EVALUATION

The methods described above were applied to speech data recorded at “normal” and “abnormal” state of the speaker. Records of both states were phonetically identical. The results presented in [5] show that only some of the LF parameters \( E_e, \omega_g, \alpha \) and \( \epsilon \) are suitable for speaker state recognition. As mentioned above the main aim was to show the dependence of analysis results on the methods for segment selection. The procedures described in the previous section (see Fig. 2 and Fig. 3) were used for the selection of segments and the results were evaluated by the following procedure:

- for both methods of segment selection ten sets were created, each set contains six segments, see Fig. 2 and Fig. 3. Recordings of one male speaker were used. Six parameters were deduced from the fact that a phoneme 40 ms long contains just six fundamental periods (thus in this case segments too) with frequency 150 Hz. In the case of more segments in the set, the selection effectiveness will decrease below admissible limits, because longer-time phonemes occur in speech less frequently.
- for each segment of the speech the glottal pulses were estimated by using an estimation of linear prediction error, by cepstral coefficients [8] or by ARMA modelling [9].
- for each set of segments the LF parameters were computed. In Fig. 6 the parameter \( \alpha \) is shown in dependence on the segment from which it was computed.
- for each set of segments the average value of the corresponding parameter (\( \mu_{\text{Normal}}, \mu_{\text{Stress}} \)) and dispersion (\( R_{\text{Normal}}, R_{\text{Stress}} \)) were computed. So, we obtained ten average values and ten dispersion values for ten sets of segments.

**Fig. 6** Values of the parameter \( \alpha \) for one segment – one set is shown, phoneme “a”, segment selection one by one in time.

In Fig. 7, the values \( \mu \) and \( R \) are shown in dependence on the set from which they were computed.

**Fig. 7** Average values of the parameter \( \alpha \) and its scatter for individual sets - ten sets, phoneme “a”, segment selection one by one in time.

### V. RESULTS

The results of the described algorithms with the final output shown in Fig. 7 are plotted in the diagrams in Fig. 8 and Fig. 9.
VI. CONCLUSION

By comparing of the diagrams in Fig. 8a and Fig. 8b, upper diagram for parameter $\alpha$, it was found that the results are better if the segments were chosen one by one in time (greater differences between the parameters for “normal” and “abnormal” speaker state). Similar conclusion can be drawn for phoneme “e” too, parameter $\alpha$. On the other hand, the parameter $\varepsilon$ is almost independent of the method of segment selection, for both analysed phonemes. Generally, it can be said that the analysis of the segments going one by one in time provides better results. The parameters are not only more different for single states than in the case of the random segment selection, but the results are also less scattered.

REFERENCES


