ADVANCED ESTIMATION OF SPEECH SIGNAL PERIODICITY

Jan Malucha

Bachelor Degree Programme (3), FEEC BUT

E-mail: xmaluc00@vutbr.cz

Supervised by: Milan Sigmund

E-mail: sigmund@feec.vutbr.cz

Abstract: This study examined the use of methods for advanced estimation of speech signal periodicity. The speech parameters of jitter, schimmer and short time period similarity were briefly explained as well as their estimation with appropriate methods, all of them integrated into a compact MATLAB program. The practical use of the program was demonstrated on the analysis of a stressed and neutral speech signal and achieved results were presented at the end of the study.

Keywords: speech signal, periodicity, stressed voice

1 INTRODUCTION

Speech signal is a very specific topic for signal data studies that can be applied to many subject fields such as robotization, security or health service. From the communication point of view, each language has a unique speech pattern which can be analysed. From the linguistic viewpoint, speech signal of the Czech language can be split into individual words that are separated from each other by intervals of silence. Furthermore, the words can be perceived as groups of letters.

If the signal is examined in the field of acoustics, we are able study the attributes of each sound in a specific time interval. In such case, it is possible to divide individual phonemes (i.e. spoken letters) into two categories: voiced and unvoiced [1]. The unvoiced phonemes are not created by vocal cords vibrations but they are based on noise excitation (e.g. phoneme /s/). On the other hand, the voiced phonemes (e.g. phoneme /a/) are created by vocal cords vibrations in a quasiperiodic process. Quasiperiodicity is a non-ideal periodic process characteristic for the presence of deviations in time behaviour which is a typical phenomenon for a real-life environment. If analysed on the elemental level, we can split speech signal of voiced phonemes into short-time period microsegments.

2 PERIODICITY

Under ideal circumstances, the term *periodicity* has a binary characteristic – the signal is either periodic or not. However, we also need to consider all impacts affecting the excitation and signal transmission. These impacts cause deviations in signal periods - more specifically, they affect the duration of individual periods and their instantaneous values. We distinguish two types of deviations that can be quantified and measured: jitter and schimmer [1]. At this point, the term *periodicity* acquires a new meaning – rather than a binary attribute, it represents the amount of approaching the state of ideal periodicity. In order to be able to examine the quasiperiodic speech signal and its periodicity properly, we need to define physical quantities that would help us to express the properties of jitter and schimmer. Jitter relates to the variation of fundamental period duration measured from cycle to cycle. It is used for the description of speech intonation - sometimes referred to as a pitch. In spoken language, the fundamental frequency ranging between 50–600 Hz is comprehended as a melody. Conversely, schimmer - perceived as deviations in instantaneous values of individual periods - can be understood as a parameter for describing the similarity of two adjacent periods. We often perceive schimmer as a quivering voice during speech.

3 PROGRAM DESCRIPTION

For automatic estimation of speech signal periodicity, a new program was created in MATLAB environment. Its main functions use methods for signal processing (including pre-processing) in short-term analysis of speech signal. The input signal is split into short segments of an optional length from 20 ms to 40 ms that are examined further as shown in Fig. 1.

To determine whether a phoneme is voiced or unvoiced, low level methods are used as follows: short time energy (STE), zero crossing rate (ZCR) and harmonic-to-noise ratio (HNR). All these methods proceed from the previously mentioned fact that any unvoiced phoneme is naturally based on noise. STE algorithm checks different values of noise short-time energy and periodic signal [1], ZCR method is based on checking the different amounts of signal crossings of zero level [2] and HNR method considers the relation between harmonic and noise signal components [3].

The fundamental frequency is then estimated in voiced speech parts only using auto-correlation function (ACF), average magnitude difference function (AMDF) and normalized cross-correlation (NCC). ACF is a widely used function to determine similarity of the examined signals from its time lag based on multiplication of shifted signals [4]. The other two methods are ACF modifications; NCC is able to follow steep changes in the input signal more accurately due to normalization [5] and in ADMF is the multiplication substituted by subtraction [6].

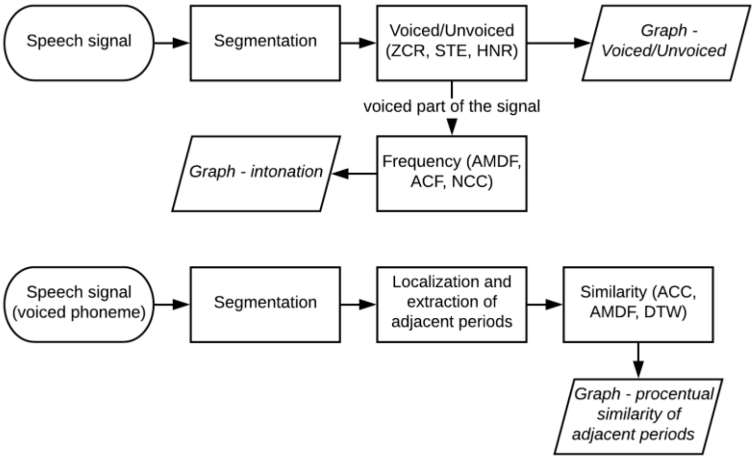
Speech signal
Segmentation
Voiced/Unvoiced (ZCR, STE, HNR)
Graph - Voiced/Unvoiced
voiced part of the signal
Frequency (AMDF, ACF, NCC)
Graph - intonation
Speech signal (voiced phoneme)
Segmentation
Localization and extraction of adjacent periods
Similarity (ACC, AMDF, DTW)
Graph - procentual similarity of adjacent periods

Figure 1: Block diagram of the MATLAB program

To be able to compare waves of adjacent periods with regard to deviations of period duration, we apply an algorithm for length adjustment that can either linearly or non-linearly reduce the length of the longer period to a length of the shorter one. The non-linear time alignment is based on dynamic time warping ACC algorithm [7]. After length adjustment, the similarity of adjacent periods is determined by ACC, AMDF or Dynamic Time Warping (DTW), which defines the similarity from the shortest Euclidean distance of compared signals.

4 PRACTICAL APPLICATION AND TESTING

The practical use of the developed program lies in a broad variety of new options when examining the speech periodicity. For example, we are able to study how much the speech signal is impacted by stress. The speech itself is created by controlling the vocal tract muscle tension. According [8], stress in speaker can affect this process and cause a loss of proper voice control by increasing the respiratory rate and vocal tract muscle tension. This is easily recognized as changes in voice fundamental frequency, speech intonation, quivering voice, speech rate and a number of pauses during the speech. We studied the stress effects via analysis of two kinds of Czech speech signals (22 kHz, 16 bits, mono, wav format). The first track was recorded during the final state examination at our faculty and the recorded voice is strongly affected by stress. The other track was recorded with a neutral voice by the same speaker. The text is identical in both cases. One male person was selected for experiments. The stressed signal is 6:25 minutes long. That is two times longer than the neutral signal which lasts for 3:27 minutes. The neutral speech signal length is sufficient for the statistical reliability of experiments [9]. The difference in the length of tracks was caused by the speaker´s frequent pausing and stuttering during the stressed speech.

The first parameter we analysed was voicing. It was expected that the unvoiced part of the speech affected by stress would be greater. The signals were split into frames of 20ms. The frames were sorted out as voiced or unvoiced by the STE algorithm and the obtained data were used to create histograms displayed in Fig. 2. As can be seen, the stressed speech track is slightly more voiced. Although the recorded track of the stressed speech is almost two times longer than the neutral speech track, it is noteworthy that the number of unvoiced segments is exactly twice as large but the number of voiced segments did not increase as much – approx. 1.38 times only. For this reason, it is possible to assume that pauses and stuttering appear far more often in stressed speech.

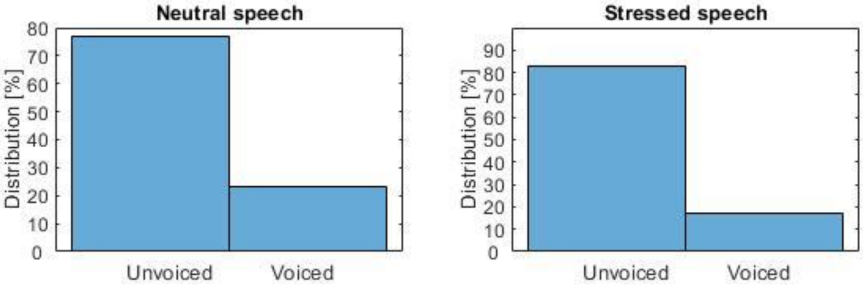

Figure 2: Histograms of voiced and unvoiced frames

The next analysed voice parameter was fundamental frequency. In general, vocal cords tension can increase due to stress and as a consequence the fundamental frequency may be raised. We examined this phenomenon using AMDF algorithm and the obtained total results are presented in the form of histograms in Fig. 3. It is apparent that fundamental frequency values of the neutral speech are concentrated mainly around 100 Hz and the higher frequencies influence the melodic accent. However, fundamental frequencies of the stressed speech are concentrated close to 127 Hz (not exceeding a total range of 70 Hz) and it completely lacks any melodic aspect. In other words, the voice pitch increased during the stressed speech and the voice became monotonic.

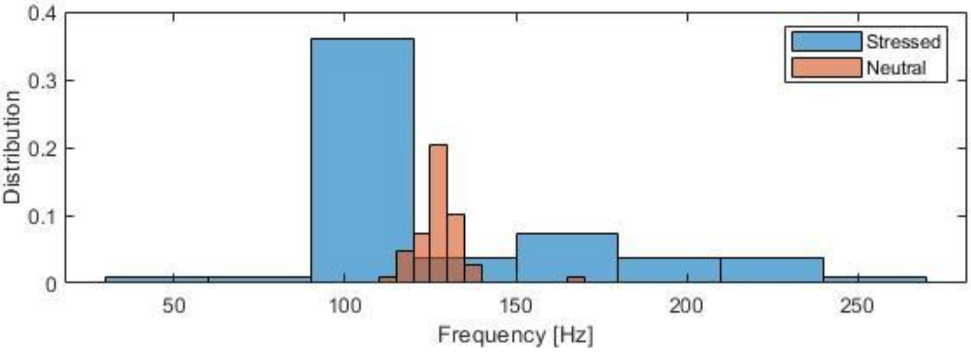

Figure 3: Histograms of voice fundamental frequencies

Schimmer was the last parameter investigated experimentally. We expected that acute stress may cause a quivering voice. In order to verify this hypothesis, several voiced phonemes /a/ of 10-period-length were extracted from the same parts of linguistic content in both speech signals. These phonemes were then examined by AMDF-based algorithm with a linear length adjustment. The similarity of all adjacent period was pairwise compared. Figure 4 illustrates that the periodicity of voiced phonemes extracted from the neutral speech is higher than the periodicity of phonemes extracted from the stressed speech.

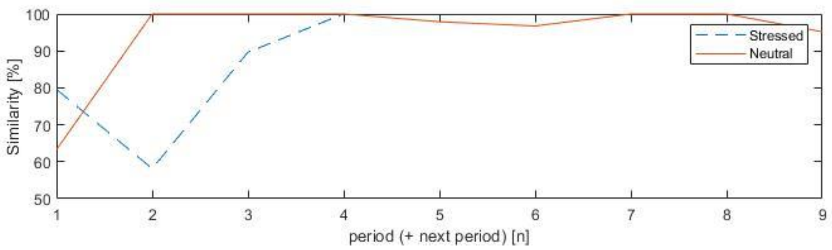

Figure 4: Diagram of adjacent period similarities for the phoneme /a/

5 CONCLUSIONS

To summarize the speech signal analysis provided by the developed program, the effect of acute stress on the speech signal were confirmed. Apart from the existing mathematical methods, the program disposes of a unique algorithm designed to split phonemes into separate periods, to adjust the length of individual periods in both linear and non-linear way as well as to create graphs of adjacent period similarities. Using the procedures, there is a variety of options for practical speech signal analysis and study of signal-impacting issues such as various voice dysfunctions or diseases. In future work, it will be useful to investigate the influence of nonstationary noise [10] and different vocal effort [11] on the accuracy of measured results.

REFERENCES

[1] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*. Prentice Hall, London, 2011.

[2] R. G. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy. In *Proc. Advanced Techniques in Computing Sciences and Software Engineering*. Springer, Dordrecht, 2010, pp. 279-282.

[3] J. P. Teixeira, C. Oliveira, and C. Lopes, Vocal acoustic analysis-jitter, shimmer and HNR parameters, *Procedia Technology*, 2013, vol. 9, pp. 1112-1122.

[4] L. R. Rabiner, On the use of autocorrelation analysis for pitch detection, *IEEE Trans. Acoust., Speech, and Signal Processing*, 1977, vol. 25, no. 1, pp. 24-33.

[5] P. Motlíček, Estimation of fundamental frequency in speech. In *Proc. 1st Conference of Czech student AES*. 2000, FEEC BUT Brno, pp. 1-6.

[6] L. Tan and M. Karnjanadecha, M. Pitch detection algorithm: autocorrelation method and AMDF. In *Proc. 3rd International Symposium on Communications and Information Technology*, 2003, vol. 2, pp. 551-556.

[7] *Dynamic Time Warping* [online]. Available: https://towardsdatascience.com/

[8] E. S. Jackson, M. Tiede, D. Beal, and D. H. Whalen, The impact of social–cognitive stress on speech variability, determinism, and stability in adults who do and do not stutter. *Journal of Speech, Language, and Hearing Research*, 2016, vol. 59, no. 6, pp. 1295-1314.

[9] M. Sigmund, Statistical analysis of fundamental frequency based features in speech under stress. *Information Technology and Control*, 2013, vol. 42, no. 3, pp. 286-291.

[10] P. Zelinka and M. Sigmund, Hierarchical classification tree modeling of nonstationary noise for robust speech recognition. *Information Technology and Control*, 2010, vol. 39, no. 3, pp. 202-210.

[11] P. Zelinka and M. Sigmund, Automatic vocal effort detection for reliable speech recognition. In *Proc. International Workshop on Machine Learning for Signal Processing*, 2010, Kittila, pp. 349-354.