ISSN 1392–124X (print), ISSN 2335–884X (online) INFORMATION TECHNOLOGY AND CONTROL, 2013, Vol. 42, No. 3

Statistical Analysis of Fundamental Frequency Based Features in Speech under Stress

Milan Sigmund

Brno University of Technology, Faculty of Electrical Engineering and Communication, Dept. of Radio Electronics, Technicka 12, 616 00 Brno, Czech Republic
e-mail: sigmund@feec.vutbr.cz

crossref http://dx.doi.org/10.5755/j01.itc.42.3.3895

Abstract. A significant part of the non-linguistic information carried in speech refers to the speaker and his/her internal state. This study investigates sixteen features based on fundamental frequency of speech F_0 in order to detect stress in speakers. The most effective features resulting from experiments are presented here. The total frequency ranges of F_0 across specific short-time speech segments created by two or three frames having stable F_0 values were evaluated as the best features for speaker-independent stress detection. F_0 contours were computed frame-by-frame using an optimized autocorrelation function. In our experiments, we used utterances spoken by 14 male speakers and taken from own database of speech under real psychological stress.

Keywords: speech signal processing; fundamental frequency; statistical evaluation.

1. Introduction

Speech presents two broad groups of information. It carries linguistic information and information about manner of speech production having no linguistic function. The fundamental frequency of speech signal (usually abbreviated as F_0) is a widely used non-linguistic speech feature which can be directly identified by human observers – it is well audible. Fundamental frequency is also one of the main factors which can distinguish the speaker's sex. Typical values of F_0 are 110 Hz for male speech and 210 Hz for female speech. Most values of F_0 among people aged 20 to 70 years lie between 80-170 Hz for men, 150-260 Hz for women and 300-500 Hz for children [1]. There are Gaussian distributions of these ranges. Usually, the mean of F_0 varies slightly during the time of day; in the morning it is lower than in the evening. Humans perceive fundamental frequency as pitch. However, perceived pitch (subjective attribute) is influenced by both sound frequency (physical attribute) and sound intensity. The human sensitivity to pitch is sharper than the sensitivity to resonance bandwidth of vocal tract [11], represented by formants.

Fundamental frequency is an important feature that characterizes the individual speakers and their emotional state. Considering the function of vocal tract, fundamental frequency is short-term determined by the rate at which vocal cords vibrate at any given time. The mean fundamental frequency characterising a speaker is determined primarily according to the membranous length of the speaker's vocal folds [14].

In West languages, i.e. in all European languages, dynamic variability of F_0 relates principally to the intonation of spoken words [8]. On the contrary, in east tonal languages such as Mandarin-Chinese, Thai, and Vietnamese, F_0 contours distinguish the meaning of words. For instance, Chinese has four relatively distinct tone types, i.e. constant, rising, falling, and falling then rising [16].

1.1. Methods for estimating fundamental frequency

By the signal theory, fundamental frequency can be seen as the lowest frequency in a harmonic series representing periodic parts of a speech signal. Because of its importance, many methods for estimating fundamental frequency have been proposed and widely studied in speech processing literature. All developed methods generally fall into four categories depending on the features' domain; i.e. time domain, frequency domain, hybrid time and frequency domain, and event detection methods.

The most obvious way to measure the F_0 value is to derive it from the speech waveform. However, accurate and reliable measurement of F_0 from the acoustic waveform alone may be in some cases quite difficult because the speech waveform varies both in period and in the detail structure of the waveform

within a period. One of the first algorithms to appear, and one of the simplest, is an algorithm that uses multiple measurements of periodicity in the signal and chooses between them to determine the voicing state and fundamental frequency. This algorithm was originally known in literature as the Gold-Rabiner's algorithm [3], and motivated many other variants based on time-domain measurements. Some basic algorithms operating in the time domain as well as in the frequency domain are described in [11]. One event-based algorithm utilizing the dyadic wavelet transform is introduced in [7]. Usually, the event-based methods attempt to determine the instant when the glottis closes and thus these detectors can accurately estimate the individual periods within a time segment since they do not assume quasi-stationarity during the measurement interval. An effective method for estimating F_0 of the vocal part in polyphonic audio signals can be found, for example, in [7]. This approach is also applicable to a speech signal with background music. Finally, an algorithm that is rarely used in real-time speech systems, but often used for research experimentation, operates on the cepstrum of the speech signal [10]. This algorithm is still popular today as an accurate method for estimating the fundamental frequency in extremely quiet laboratory recording conditions. The details of individual algorithms are beyond the scope of this paper. A good overview of fundamental frequency determination is given, for instance, in the monograph [5] and in [6].

In general, all of the proposed methods have their limitations, and no presently available algorithm can be expected to give perfect F_0 values across a wide range of applications and operating environments. The measured raw values of F_0 usually need to be post-processed in order to eliminate isolated errors. Post-processing algorithms involve smoothing the derived F_0 contour, rejecting too short voiced or unvoiced segments, rejecting low-energy voiced segments, etc. For instance, a set of post-processing techniques applied to five F_0 determination algorithms is introduced and its performance is evaluated in [15].

1.2. Outline

In Section 2, our way of speech processing and F_0 estimation will be introduced. Experimental setups are reported in Section 3 which is divided into two subsections. The first subsection describes speech materials used in our experiments. In the second subsection we present the main findings on the effect of psychological stress on statistical parameters of several F_0 characteristics. Finally, Section 4 briefly concludes the paper and gives some suggestions for future work.

2. Algorithm used for determining F_0

In our experiments, the fundamental frequency F_0 was estimated frame-by-frame on the basis of a modified autocorrelation function. In this algorithm, voicing and fundamental frequency are computed simultaneously using the high peaks of a speech signal only while other signal samples on middle and low levels are suppressed [11].

Firstly, a clipping level CL was set in each frame m as a fixed percentage of the smaller value of two maximum signal amplitudes measured in the previous frame m–1 and in the following frame m+1. A clipped signal $\tilde{s}(n)$ obtained from the speech signal $s(n)$ results, after amplitude normalization, in three different values only: $\tilde{s}(n) = +1$ if $s(n) > CL$, $\tilde{s}(n) = -1$ if $s(n) < -CL$, and $\tilde{s}(n) = 0$ otherwise. Then, the short-time autocorrelation of the clipped signal was estimated as

$$R(k) = \sum_{n=1}^{N-k} \tilde{s}(n)\,\tilde{s}(n+k), \qquad (1)$$

where N denotes the length of the speech frame, and k lies in the interval $0 \le k \le N-1$. Figure 1 illustrates two typical autocorrelation functions obtained from both a voiced and unvoiced speech frame.

Furthermore, the autocorrelation wave was processed. The highest peak in the autocorrelation function, except for $R(0)$, must be located (as marked with a vertical dashed line at k_P in Fig. 1) and then its value is compared to a fixed threshold derived from the $R(0)$ value. If the peak exceeds the threshold, the frame is classed as voiced else as unvoiced:

$$R(k_P) \ge \alpha\, R(0) \quad \approx \quad \text{voiced}, \qquad (2)$$

$$R(k_P) < \alpha\, R(0) \quad \approx \quad \text{unvoiced}. \qquad (3)$$

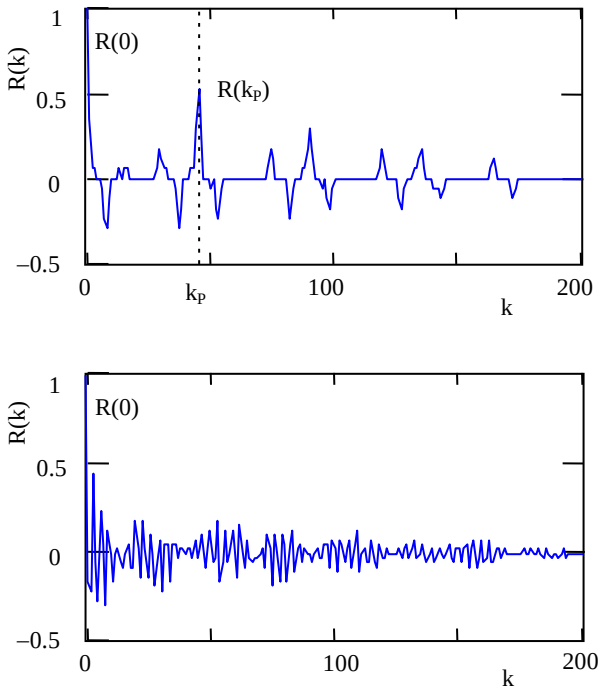

Figure 1. Autocorrelation function of clipped voiced speech (top) and unvoiced speech (bottom)

In the case of a voiced frame, the fundamental frequency is defined by the position of the highest peak k_P and by the sampling frequency of the speech signal f_{sam}

$$F_0 = \frac{f_{sam}}{k_P}. \quad (4)$$

For unvoiced speech, the fundamental frequency is undefined and set $F_0 \equiv 0$ by convention. A more detailed description of this algorithm also including a discussion can be found in [11].

In order to optimize the used algorithm, several reasonable combinations of numeric values for the clipping level *CL* and for the threshold α were examined on the basis of normal speech. Generally, the autocorrelation threshold defines how strong the selection of voiced frames is, while the clipping level gives accurate indication of the F_0 values. For each combination of *CL* and α, the percentage of frames *v* classified as voiced and error rate *e* for F_0 estimation were investigated. Here, the error rate was calculated by a simple approach considering gross errors only represented by isolated F_0 estimates outlying from a frequency band defined by interval ±50 Hz around the mean of F_0. Table 1 summarizes the averaged results obtained from speech signal spoken by five speakers.

Table 1. Evaluation of the F_0 estimation algorithm for more numeric combinations of parameters *CL* and α

CL	α = 0.3		α = 0.4		α = 0.5	
	e [%]	v [%]	e [%]	v [%]	e [%]	v [%]
0.5	3	31	1	27	0	19
0.6	5	38	2	32	0	25
0.7	7	41	4	35	1	26
0.8	8	42	5	33	2	20
0.9	13	30	9	21	4	11

In the later experiments performed in this study, fixed values of *CL*=0.6 and α=0.4 were used. The choice for these values can be seen as a compromise between low error rate and high efficiency of the selection of voiced frames. However, in some cases, extraneous peaks can appear in the autocorrelation function which decreases the accuracy of F_0 estimation. To obtain the true values of F_0 for statistical analysis, the search range of k_P was limited to $32 \le k_P \le 100$, corresponding to the F_0 range of 80-250 Hz (for male speech in our case). Additionally, irregularities in F_0 contour doubling and halving F_0 were eliminated. Using the described algorithm, Figure 2 shows an example of a detailed F_0 contour of the Czech word *večer* (meaning *evening* in English) spoken by a male speaker.

It should be noted that there exists no best method to estimate fundamental frequency. We applied the above described algorithm because it is robust against noise, produces good estimates of the fundamental frequency, requires only a small number of standard arithmetic computations, and thus it can be easily implemented in digital hardware.

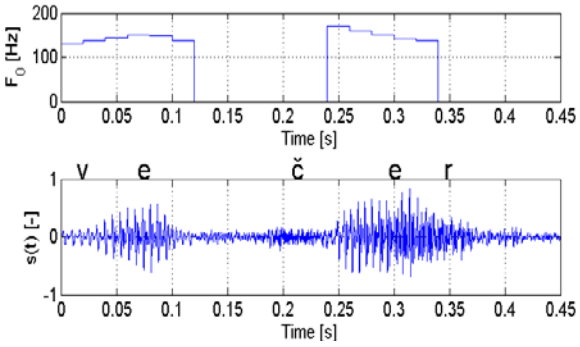

Figure 2. Czech word "večer" aligned with its speech waveform and corresponding F_0 contour

3. Experimental Results

3.1. Speech data

A suitable corpus of speech data is a very important prerequisite for effective speech research. Although a range of databases with speech under stress exist, they cover very different types of speech data and are only partly useful for research into stress. Available databases were created mostly by recordings under simulated stress. Some information about existing English and German databases appropriate for research into speech under stress may be found in [13].

In our experiments we used Czech speech material from the database Exam Stress collected under real psychological stress. In this database, stress was induced by the final examinations at our university held in oral form in front of a board of examiners. The speakers were male pre-graduate students (Czech natives). The speech data contains paired neutral speech and stressed speech spoken by all speakers. The examinations at a university can be considered as a medium stress condition influencing the individual students in different degrees. The database Exam Stress was recorded with the task of stress identification in mind.

In our experiments, the time series of fundamental frequency was estimated on a frame basis in both neutral and stressed speech in fourteen speakers. The speech signal from the database Exam Stress (22 kHz, 16 bits) was resampled at 8 kHz and short-time segmented by a rectangular window (20 ms) without overlapping.

3.2. Statistical results

The fundamental frequency of voice F_0 and some of its derivatives were investigated independent of pronounced words using statistical parameters. The first measurement was focused on the estimation of length of speech signal needed for calculation of

reliable basic statistical parameters of F_0. Experimental results show that a data set of approximately 2500 values (i.e., 2500 voiced speech frames) satisfies statistical reliability. Figure 3 illustrates the development of cumulative values of mean $\mu(F_0)$ and standard deviation $\sigma(F_0)$ calculated from two different parts of a male speech.

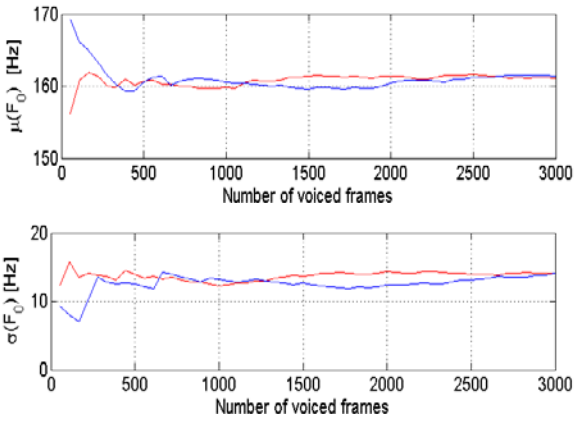

Figure 3. Cumulative values of mean μ (top) and standard deviation σ (bottom) of the fundamental frequency obtained from two different sets of 10 to 3000 voiced speech frames

A series of experiments was conducted to study the effects of stress on fundamental frequency. As a first step, F_0 contour was calculated for every speech signal. Then, parts with local short-time stable F_0 were investigated in each F_0 contour. In this step, "twins" and "triples" were detected which represent short chains of two or three adjacent frames of 20 ms having constant F_0 (taking into account F_0 values rounded to the nearest whole number). Chains of four and more adjacent frames containing constant F_0 occur very rarely (less than 0.5% of voiced frames). Thus, they were not considered in our analysis. Figure 4 illustrates the individual histograms of F_0, *twins* and *triples* for one male speaker measured in both states; neutral and under stress.

In the long-term analysis of F_0 data for detection of stress, the following 4 parameters of distribution were used in our experiments:

1. Arithmetic mean μ;
2. Standard deviation σ;
3. Skewness (3rd order moment) *skew*;
4. Kurtosis (4th order moment) *kurt*.

Skewness is a measure of the asymmetry of the data around the mean. Kurtosis measures the degree of peakedness of a distribution and it is zero only for Gaussian distribution. Applying the 4 parameters on distributions of all three features, i.e. F_0, *twins* and *triples*, give 12 features. In addition, frequency range and score of *twins* and *triples* were taken into account. All together 16 features were computed for every speech signal. Frequency range is given as a difference between the maximum and minimum frequencies at which *twins* or *triples* occur. Score (in percentage) is defined as the ratio of the number of *twins* or *triples* to the number of all voiced frames in an utterance. Table 2 shows a numerical example of all these features for the same speaker as in Fig. 5.

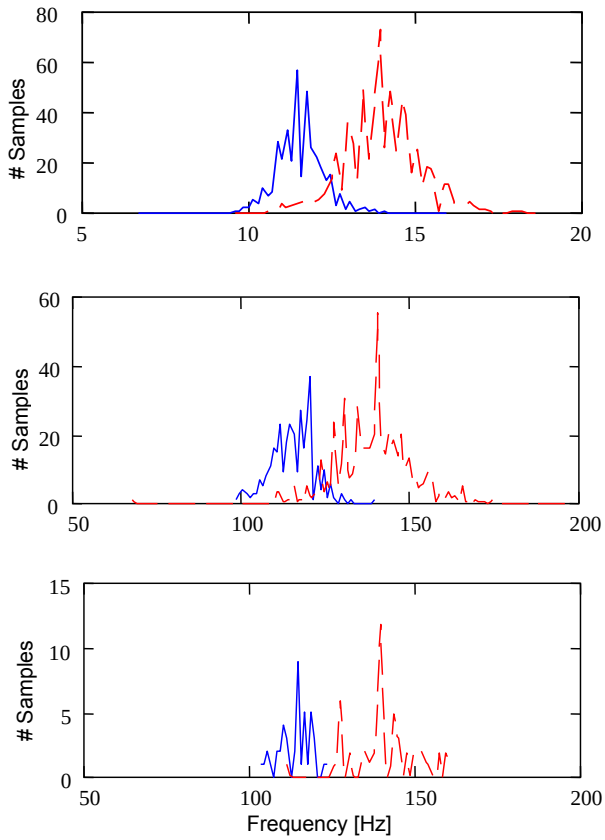

Figure 4. Individual histograms of F_0 (top), *twins* (middle) and *triples* (bottom) for the speaker Ka. Solid line graphs stand for neutral speech, dashed line graphs stand for stressed speech

Table 2. Statistical results for the speaker Ka

Feature	Neutral speech	Stressed speech
$\mu(F_0)$	115.3	140.1
$\sigma(F_0)$	7.3	11.5
skew(F_0)	0.039	0.126
kurt(F_0)	4.125	0.992
μ(*twins*)	114.9	139.0
σ(*twins*)	6.6	11.8
skew(*twins*)	0.030	–0.396
kurt(*twins*)	0.352	3.857
range(*twins*)	42	116
score(*twins*)	8.5	8.0
μ(*triples*)	113.8	141.7
σ(*triples*)	5.2	9.9
skew(*triples*)	–0.302	–0.190
kurt(*triples*)	–0.584	0.486
range(*triples*)	21	51
score(*triples*)	1.14	1.02

The numerical movements of each feature due to stress can be classified into three broad categories: rising, falling and no changing. Table 3 summarizes the movements of all features measured in all 14 speakers. In this case, weak changes of feature values (less than ±5% of the value in neutral speech) were considered as in the category "No change".

Table 3. Summary of up and down movements of feature values in stressed speech

Feature	Number of speakers		
	Rising	No change	Falling
$\mu(F_0)$	13	1	0
$\sigma(F_0)$	12	0	2
skew(F_0)	12	0	2
kurt(F_0)	6	0	8
μ(*twins*)	12	2	0
σ(*twins*)	10	2	2
skew(*twins*)	8	0	6
kurt(*twins*)	5	0	9
range(*twins*)	13	1	0
score(*twins*)	7	2	5
μ(*triples*)	13	1	0
σ(*triples*)	11	0	3
skew(*triples*)	5	0	9
kurt(*triples*)	11	0	3
range(*triples*)	13	1	0
score(*triples*)	3	3	8

Here it is evident that the used features contain different amounts of information relevant to stress detection. To rate the discriminative power of each feature *x*, the individual features were evaluated by means of two criteria. The first criterion *Q*, introduced in [9] as the quality metric Q3, is based on the ratio of intra/inter class nearest neighbour distances. In the case of two classes, i.e. two speaker's state, it becomes the simplified form as

$$Q(x) = \frac{\min\limits_{i,j,i\neq j} d(x_i^{\mathrm{n}}, x_j^{\mathrm{n}}) + \min\limits_{i,j,i\neq j} d(x_i^{\mathrm{s}}, x_j^{\mathrm{s}})}{2 * \min\limits_{i,j} d(x_i^{\mathrm{n}}, x_j^{\mathrm{s}}) + \varepsilon}, \qquad (5)$$

where d (.) denotes the squared Euclidean distance and ε is a small constant. The second criterion used, the discrimination factor *DF*, is defined as follows:

$$DF(x) = \frac{\left| \sum\limits_{k} \Delta x_k^{\mathrm{pos}} - \sum\limits_{l} \Delta x_l^{\mathrm{neg}} \right|}{\left| \mathrm{average}\,\{x_1^{\mathrm{n}}, ..., x_J^{\mathrm{n}}, x_1^{\mathrm{s}}, ..., x_J^{\mathrm{s}}\} \right|}. \qquad (6)$$

In equation (6), Δx stands for inter-state differences of the feature *x* separated into two series:

$$\Delta x_k^{\mathrm{pos}} = x_j^{\mathrm{s}} - x_j^{\mathrm{n}} \quad \text{if} \quad x_j^{\mathrm{s}} - x_j^{\mathrm{n}} \geq 0, \qquad (7a)$$

$$\Delta x_l^{\mathrm{neg}} = \left| x_j^{\mathrm{s}} - x_j^{\mathrm{n}} \right| \quad \text{if} \quad x_j^{\mathrm{s}} - x_j^{\mathrm{n}} < 0, \qquad (7b)$$

where k=1,2,… and l=1,2,... denote running indices for positive and negative differences, respectively. In equations (5) to (7), x_j^n is the j-th sample of feature x from the class "neutral speech" and x_i^s is the i-th sample of feature x from the class "stressed speech", $1 \leq i \leq J$, $1 \leq j \leq J$, and J=14 is the number of speakers. When a feature differs significantly from the neutral speech to the stressed speech, it is expected that the criterion *Q* gives low value. Conversely, low value of *DF* means bad discriminative power.

Both criteria *Q* and *DF* were applied separately for each feature listed in Table 3 across all speakers. First, all features were ranked in two lists, once by *Q* values and once by *DF* values, in ascending order. Then, the final order of features was created according to the average rank computed from feature positions in both individual rank lists. In this way, Table 4 shows the ranked features with their numeric values of *Q* and *DF*. Note that *Q* and *DF* have different numerical scale. Generally, these criteria can be applied as a fast and efficient feature pre-selection approach.

Table 4. Ranked features across all speakers in terms of *Q* and *DF*

Rank	Feature	Q	DF
1	*range(triples)*	0.5	3.8
2	*range(twins)*	0.2	3.1
3	*skew(F_0)*	14.1	4.0
4	*μ(triples)*	9.4	1.9
5	*$\mu(F_0)$*	12.2	1.6
6	*σ(triples)*	2.1	1.3
7-8	*skew(twins)*	44.6	3.6
7-8	*score(twins)*	0.8	0.06
9	*kurt(F_0)*	15.3	1.8
10	*kurt(triples)*	26.4	1.9
11	*μ(twins)*	14.7	1.5
12	*kurt(twins)*	9.5	1.3
13	*score(triples)*	7.5	0.5
14	*$\sigma(F_0)$*	25.1	1.4
15	*σ(twins)*	120	1.4
16	*skew(triples)*	109	1.2

The results show that a majority of speakers produce speech with a higher F_0 when speaking under stress (see Table 3). This effect confirms the findings in previous studies [4], [12]. On the other hand, an increase of F_0 alone may also reflect other factors influencing speech, for instance alcoholic intoxication [2] or the well-known Lombard reflex. We have measured a significant increase of mean values not only for single F_0 samples, but also for *twins* and *triples*. However, the most effective features for speaker-independent stress detection according to the criteria *Q* and *DF* seem to be the frequency range of *twins* and *triples* (see Table 4) which extend when speaking under stress.

4. Conclusion and future work

It is generally accepted that changes in the fundamental frequency of voice reflect a speaker's emotional state. While emotion researchers reported mostly on average F_0 only, we investigated the influence of exam stress on the distributions of F_0 and multiple frames having constant F_0, i.e. *twins* and *triples*. A set of 16 features based on F_0 was measured and evaluated for the capability of detecting stress. The new proposed features, *twins* and *triples* give promising results for further research. All experiments were carried out on recordings containing speech data in two classes only: neutral speech and stressed speech.

In future research, it will be necessary to enlarge our database by adding recordings with other types of emotions that can affect speech features similarly like stress. Furthermore, it will be useful to analyse the described speech features in adverse acoustic conditions using nonstationary ambient noise modeling [17]. The goal of this research is the development of algorithms for automatic detection of true stress during speech dialogue. In this way, F_0 might act as a robust feature applicable to a remote psychological check of humans operating in very responsible work places such as air traffic control, chief of command in military crisis situations, etc.

Acknowledgements

The research was supported by the WICOMT project; registration number CZ.1.07/2.3.00/20.0007 financed from the operational program Education for Competitiveness. The support of the project MOBYS financed from BUT Brno is also gratefully acknowledged.

References

[1] **R. J. Baken, R. F. Orlikoff.** *Clinical Measurement of Speech and Voice*. San Diego: Singular Publishing, 2000.

[2] **B. Baumeister, Ch. Heinrich, F. Schiel.** The influence of alcoholic intoxication on the fundamental frequency of female and male speakers. *Journal of the Acoustical Society of America*, 2012, Vol. 132, No. 1, 442-451.

[3] **B. Gold, L. R. Rabiner.** Parallel processing techniques for estimating pitch periods of speech in the time domain. *Journal of the Acoustical Society of America*, 1969, Vol. 46, No. 2, 442-448.

[4] **J. H. Hansen, S. E. Ghazale.** Getting started with SUSAS. In: *Proceedings of the 8th European Conference on Speech Communication and Technology*, Rhodes, 1997, pp. 1743-1746.

[5] **W. J. Hess.** *Pitch determination of speech signals – algorithms and devices*. New York: Springer-Verlag, 1983.

[6] **W. J. Hess.** Pitch and voicing determination. In Furui, S., Sondhi, M. (eds.). *Advances in Speech Signal Processing*. New York: Marcel Dekker, 1992, pp. 3-48.

[7] **S. Kadambe, G. F. Boudreaux.** Application of the wavelet transform for pitch detection of speech signals. *IEEE Transactions on Information Theory*, 1992, Vol. 38, No. 2, 917-924.

[8] **R. D. Ladd.** *Intonational Phonology*. Cambridge: Cambridge University Press, 1996.

[9] **R. Lileikyte, L. Telksnys.** Quality estimation of speech recognition features for dynamic time warping classifier. *Information Technology and Control*, 2012, Vol. 41, No. 3, 268-273. http://dx.doi.org/10.5755/j01.itc.41.2.914

[10] **A. M. Noll.** Cepstrum pitch determination. *Journal of the Acoustic Society of America*, 1967, Vol. 41, No. 2, 293-309.

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

[12] **M. Sigmund, T. Dostal.** Analysis of emotional stress in speech. In: *Proceedings of International Conference on Artificial Intelligence and Applications*, Innsbruck, 2004, pp. 317-322.

[13] **M. Sigmund.** Influence of psychological stress on formant structure of vowels. *Electronics and Electrical Engineering*, 2012, Vol. 18, No. 10, 45-48.

[14] **I. R. Titze.** Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America*, 1989, Vol. 85, No. 4, 1699-1707.

[15] **P. Veprek, M. Scordilis.** Analysis, enhancement and evaluation of five pitch determination techniques. *Speech Communication*, 2002, Vol. 37, No. 3, 249-270.

[16] **Y. Wu, K. Hemmi, K. Inoue.** A tone recognition of polysyllabic Chinese words using an approximation model of four tone pitch patterns. In: *Proceedings of International Conference on Industrial Electronics, Control and Instrumentation*, Kobe, 1991, pp. 2115-2119.

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

Received March 2013.