Electrical, Control and Communication Engineering
doi: 10.2478/ecce-2014-0009
2014 / 5

Improving Speech Recognition Rate through Analysis Parameters

Deividas Eringis (*PhD Student, Vilnius University Institute of Mathematics and Informatics*),
Gintautas Tamulevičius (*Researcher, Vilnius University Institute of Mathematics and Informatics*)

***Abstract* – Speech signal is redundant and non-stationary by nature. Because of vocal tract inertness these variations are not very rapid and the signal can be considered as stationary in short segments. It is presumed that in short-time magnitude spectrum the most distinct information of speech is contained. This is the main reason for speech signal analysis in frame-by-frame manner. The analyzed speech signal is segmented into overlapping segments (so-called frames) for this purpose. Segments of 15–25 ms with the overlap of 10–15 ms are used usually.**

In this paper we present results of our investigation of analysis window length and frame shift influence on speech recognition rate. We have analyzed three different cepstral analysis approaches for this purpose: mel frequency cepstral analysis (MFCC), linear prediction cepstral analysis (LPCC) and perceptual linear prediction cepstral analysis (PLPC). The highest speech recognition rate was obtained using 10 ms length analysis window with the frame shift varying from 7.5 to 10 ms (regardless of analysis type). The highest increase of recognition rate was 2.5 %.

***Keywords* – Computers and information processing; Speech analysis; Speech recognition; Speech enhancement.**

I. Introduction

Speech analysis is the main stage of the speech recognition process. The extracted data (so-called features) with high discriminating power should condition fast and reliable speech recognition process.

The essential purpose of these features is to carry specific information about linguistic content of the speech utterance. In ideal case these features should not contain information about the speaker: its gender, speaking style, age, physical condition, etc. This is not true for real-world recognition tasks and the selection of noise and speaker robust features is the main problem in speech recognition systems. A tremendous amount of research has been made in search of appropriate feature extraction and enhancing techniques in order to obtain reliable and noise robust speech recognition [1], [2] and [3]. Various static and dynamic feature systems were proposed for speech recognition: time-scale (like frame energy, zero-crossing-rate), frequency-scale features (like formant based, spectral pairs, etc.), various cepstral domain and human hearing modeling features.

The first step of any feature extraction process is the segmentation of speech signal into analysis frames – overlapping signal segments. This allows us to represent non-stationary speech signal by nature as stationary segments of the signal. Frames with constant length of 15–25 ms and overlap of 10–15 ms are usually used.

Despite these parameter values are declared as effective [3], [4] and [5] for speech analysis task, some criticism should be given at this point. The analysis frame length of 15–25 ms has been used for more than 30 years. The sampling rate of the speech records was enlarged from 6 kHz [6] up to 44.1 kHz during this period. Therefore, the rate of analyzed signal data was increased by a few times despite the same frame window length is used. As we know higher data order means higher spectral analysis resolution. And higher resolution does not mean higher speech recognition rate. In some cases additional spectral information can represent the undesirable reference to individual speaker attributes. Thus analysis parameter values should be revised considering modern speech signal analysis approaches.

In this paper we will present our investigation of analysis parameters manipulation as the recognition rate improving technique. Our goal is to evaluate the effect of analysis parameters on speech recognition rate (the analysis order will not be investigated). Finding optimal (in some sense) analysis parameters values would enable us to improve speech recognition rate without any substantial modification of speech recognition system.

This paper is organized as follows. Section II provides comparative review of feature extraction methods for speech recognition. Section III describes speech analysis parameters and their impact on analysis results. Section IV presents experimental process and results. Experimental results are presented and discussed in Section V.

II. Feature Extraction

Mel frequency cepstral (MFCC) analysis and perceptual linear prediction (PLP) analysis are the most widely used feature extraction approaches in modern speech recognition systems [7] – [10].

MFCC analysis is based on filter-bank which is considered as the model of the speech processing in human auditory system. The filter-bank is implemented as the set of triangular shaped band-pass filters arranged in non-uniform frequency scale. The mel scale and bark scale are widely used for filter-bank arrangement. Mel-frequency cepstral coefficients are calculated [5]:

$$C_{MFCC}(i)=\sum_{k=1}^{N} X_k \cos\left[i\left(k-\frac{1}{2}\right)\frac{\pi}{N}\right],\quad i=1,2,\quad \ldots\quad ,M\,, \qquad (1)$$

where i – analysis order, N – length of analyzed signal.

Regardless of MFCC wide usage [7] – [10] in speech recognition task, these features lack the ability to capture non-linearities which are related to patterns in speech waveforms [11]. Authors [11] proposed Gamma-chirp filter-bank for frequency analysis thus simulating the non-linearity of a human hearing system. For non-linear speech processing there was used the Dyn operator. This operator simulates signal processing in human auditory system. Researchers tried to optimize speech recognition by altering mel-cepstrum [12]. Characteristics of filters and the order of analysis were modified to find out whether that has any influence on recognition rate.

Various MFCC analysis enhancement techniques were proposed for robust speech recognition. The following techniques can be named: multitaper method based on averaging of several windows (tapers) in frequency domain, RASTA filtering, feature warping, normalization of cepstral coefficients.

Perceptual linear predictive cepstral analysis is based on usage of all-pole model for simulation of processing in auditory system spectrum [13]. First of all speech short-time power spectrum is calculated. Afterwards the critical-band (bark) analysis, i.e. transformation from linear frequency scale spectrum to bark frequency is performed:

$$B(f) = 6\ln\left[\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right], \quad (2)$$

where f is linear frequency in Hertz.

The equal-loudness pre-emphasis and loudness compression follows next. This models the intensity of the perceived speech and its loudness in hearing system. The approximation of the power spectrum by an all-pole model using autocorrelation method and cepstral analysis follows finally [14].

Linear prediction (LPC) based analysis approaches seek to model the speech generation process thus extracting the appropriate speech data. The LPC model represents speech production process as "excitation source-system" (glottis–lips) scheme where the source signal (vibration of the vocal chords) and the system (vocal tract) can be modeled apart. This allows us to model the vocal tract using linear filter thus rejecting the speaker specific information. More detailed LPC and LPC derivative analysis can be found in [14].

Using discrete wavelet transformation temporal information is extracted by shifting and rescaling main wavelet [15], [16]. For this purpose speech signal is analyzed using particular resolutions at different frequencies. Discrete wavelet transform extracts not only frequency information of speech signal but also linguistically relevant temporal information, such as [16] and [17]:

- vocal tract transfer function information, which contains amplitude and duration references of voicing and articulation, vowel length;
- information about periodicity and its variation;
- fluctuations at higher frequencies.

Some modern speech analysis approaches for feature extraction are proposed. Minimum Variance Distortionless Response (MVDR) spectrum estimation is used to estimate speech envelope (which represents vocal tract transfer function) and is declared as more robust to additive noise [18]. Another approach normalizes the ratio of arithmetic and geometric means of power spectrum coefficients, which ought to be applicable in the situations when the speech is corrupted by noise heavily [19].

The analysis parameters (the length and the shift of the analysis frame) are chosen "traditionally": the length of analysis frame is 15–25 ms usually with the shift of 10-15 ms. The order of static feature analysis is set to 12–15 plus various differential feature forms [4], [20] – [24], [3] and [5]. These values of analysis parameters are used without any argumentation or questioning, though it is evident that in some cases revised analysis parameters will have influence on speech recognition rate.

In the next section we analyze the effect of the analysis window length and shift size on speech recognition performance.

III. Speech Analysis Parameters

Since human speech signal is non-stationary this state of the signal can be eliminated using short-time analysis. The short-time spectral analysis is performed using short-time Fourier transform:

$$S_n\left(e^{j\omega_i}\right) = \sum_m s(m)w(n-m)e^{-j\omega_i m}, \quad (3)$$

where $S_n\left(e^{j\omega_i}\right)$ is Fourier transform of the windowed signal $s(m)w(n_0\text{-}m)$, n here represents analysis window shift in samples.

The frequency resolution of spectral analysis Δf is related to the length of the analyzed signal N and the sampling frequency F_s. Considering the length of the analyzed signal is determined by the length of the analysis window t_w we can get

$$\Delta f = \frac{F_S}{N} = \frac{F_S}{t_w \cdot F_S} = \frac{1}{t_w}. \quad (4)$$

The longer analysis window is used the higher frequency resolution and the lower time resolution is obtained. Thus the analysis window of 20 ms gives us frequency resolution of 50 Hz. If we shorten the analysis window down to 10 ms we will get higher time resolution but the frequency resolution will be reduced to 100 Hz and this can burden the analysis process of low frequency pitch-based speech and its harmonics. Therefore, the spectral analysis result is affected by the window length inevitably. Various solutions are proposed for selection of analysis window length.

Rabiner and Schafer [25] have acknowledged, that if analysis window is too large, analysis cannot reflect variation of characteristics of speech signal adequately. This suggests that there is no single one and truthful window length value for all cases, because pitch period varies within speakers: for female or children it is higher, for male – lower. Pitch period value fluctuates from 3.3 to 16.6 ms (pitch frequency varies

from 60 to 300 Hz). If we expand the length of the analysis window, the analyzed speech segment may become non-stationary. If analysis frame length will be shortened too much, some signal characteristics will be lost. Besides, if we shorten analysis window until it becomes shorter than pitch period (2–3 ms) we may miss to register pitch peaks.

Fig. 1 presents short-time magnitude spectra of Lithuanian phoneme [a] (taken from Lithuanian word "hello"). The spectrum was estimated using different length analysis windows: 10, 20, 50, 100, 200 ms, respectively. We can see that longer analysis frame gives higher resolution, thus we can analyze separate pitch harmonics. However, if the frame size is too high, transform gives us redundant frequency resolution which can give the effect of noise. This would be crucial for correct speech recognition rate.

It is proposed to use analysis window with the length of two or three pitch periods, i.e. 5–20 ms for high pitch speakers and 20–50 ms for low pitch speakers [5], [26] and [27] .

Other researchers claim that longer-duration windows (50–100 ms) overpass shorter ones in the mean of noise compensation [19].

In all aforementioned cases the window length is fixed and constant (for different speakers). This gives simpler but non-optimal analysis process. If fixed-size window is applied the analyzed signal segment may consist of quasi-periodic and non-periodic parts of speech and analysis of such segment would result in inadequate frequency data. This would be the source of uncertainty in speech recognition process.

There are proposed alternative analysis window implementations to improve speech recognition rate.

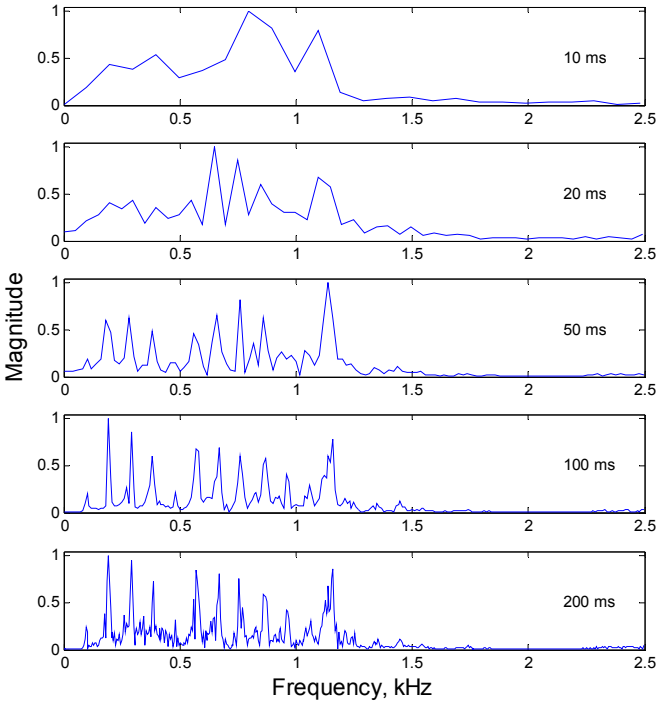

Fig. 1. Short-time Fourier transform, using various analysis window lengths. It is apparent that using wider frame gives better spectral resolution.

The first one is to use pitch period synchronized window length and shift size of analysis window [20] and [28]. Speech signal is segmented using the size of pitch contour of the signal and analysis window length and frame shift sizes are chosen as multiples of pitch period to minimize spectral distortion. In this case we get fixed but speaker-dependent analysis parameter values. Two problem sources should be pointed for pitch synchronized analysis parameters.

First of all there is not the only pitch determination approach giving reliable results. There are various determination approaches proposed giving slightly different results. And even minor inaccuracy of pitch period determination can result in spectral distortion of extracted features. Secondly, the pitch varies during the speech (especially if the speech is emotionally aroused). In this case pitch-related and fixed frame length based analysis can give distorted representation of the speech.

Therefore, variable-length frame based analysis technique is proposed [21] and [29]. Pseudo-pitch synchronous feature extraction aligns the length of analyzed signal segment to its natural cycle thus avoiding pitch period truncation [30]. The frame size can depend on speaking rate, particular sounds also [29].

Another important aspect of analysis is the degree of overlapping or analysis window (frame) shift. The size of window shift determines the particularity of information about speech dynamics. The lower frame shift value we use, the more details we can retrieve about speech dynamics. However, such analysis will take longer and it does not necessarily mean higher speech recognition rate.

The overlap is selected equal to half or one third of the analysis window length usually [20] – [22]. However, some original ideas for setting the frame shift (or the frame rate) are proposed.

The frame rate selection approach based on a posteriori signal-to-noise ratio is proposed in [31]. The approach is capable of assigning higher frame rate to a rapid changing state of speech and lower frame rate to more steady-state of speech. In [21] the size of overlap is set to half of analysis window and recursively is reduced till minimum LPC residual error is found. Another proposal is to vary frame shift according to phonetic information [.

The lengths of the speech signal and analysis window, frame shift are related:

$$K = \frac{N - L}{\Delta L} + 1, \tag{5}$$

where K – the number of extracted frames in speech signal, N – the length of the signal, L – analysis window length, ΔL – the shift of analysis window.

Usually we have $L \ll N$, thus the increase or decrease of analysis window length will not necessarily influence on the number of extracted frames to be analyzed. Therefore, the main criterion of analysis window length selection is the frequency resolution.

The frame shift determines the number of extracted frames. The smaller is the window shift, the more signal frames are extracted, the longer it will take to analyze the signal. The size of frame shift ΔL is directly related with the number of analysis frames K, e.g. if we increase frame shift twice, the number of frame we get will be twice smaller, thus analysis time should diminish twice also. This is important in case of speech recognition on the devices with limited calculation resources. Frame shift control can save calculation time and power.

We state that by varying analysis window length and shift we can improve speech recognition process. The improvement can be evidenced by higher recognition rate or faster recognition process.

IV. Experimental Setup and Results

We will investigate the isolated word recognition rate dependence on analysis window length and shift size.

For this purpose we used recordings of Lithuanian isolated words pronounced by 8 speakers (4 women and 4 men). Utterances were captured using 8,000 Hz sampling rate, mono channel and 16 bit quantization. The training set contained 800 patterns (8 speakers x 100 words x 1 pronunciation). 800 patterns were intended for testing.

We used Dynamic Time Warping (DTW) based recognizer. It is a pattern comparison based recognition approach allowing simple and effective modeling the recognition of isolated words.

Three different cepstral analysis techniques were used for feature extraction: mel-frequency cepstral coefficients (MFCC), linear predictive cepstral coefficients (LPCC) and perceptual linear predictive cepstral coefficients (PLPC). This enabled us to use the same Euclidean distance calculation scheme for all analysis techniques. The recognizer was implemented in MATLAB environment.

The goal of the first experiment was to determine the window length giving the highest recognition rate. We varied the analysis window length from 10 to 30 ms (with the increase step of 2.5 ms) for all analysis types for this purpose. The results of the first experiment are given in Fig. 2.

As we can see the highest average recognition rate was obtained using 10 ms window length for all analysis types. This length is twice shorter than widely used value of 20–25 ms. The recognition rate using 10 ms frame size was increased by up 1.2 % in comparison with 20 ms frame size case (for PLP analysis).

Analyzing the average recognition rates for female and male speakers separately we noticed that the "optimal" window (giving the highest recognition rate) for female speakers was slightly shorter than for male speakers and ranged from 10 ms to 12.5 ms for all analysis types. The "optimal" length of analysis window for male speakers was more inconsistent and ranged from 10 ms (in the case of PLPC analysis) to 17.5 ms (for LPCC analysis).

The increase of window length reduced the average recognition rate. However, the window length increase up to 15 ms in PLP analysis case and increase up to 20 ms in case of LPCC and MFCC analyses resulted in decrease of recognition rate by less than 1 %. This decrease can be considered as negligible thus acceptable if needed.

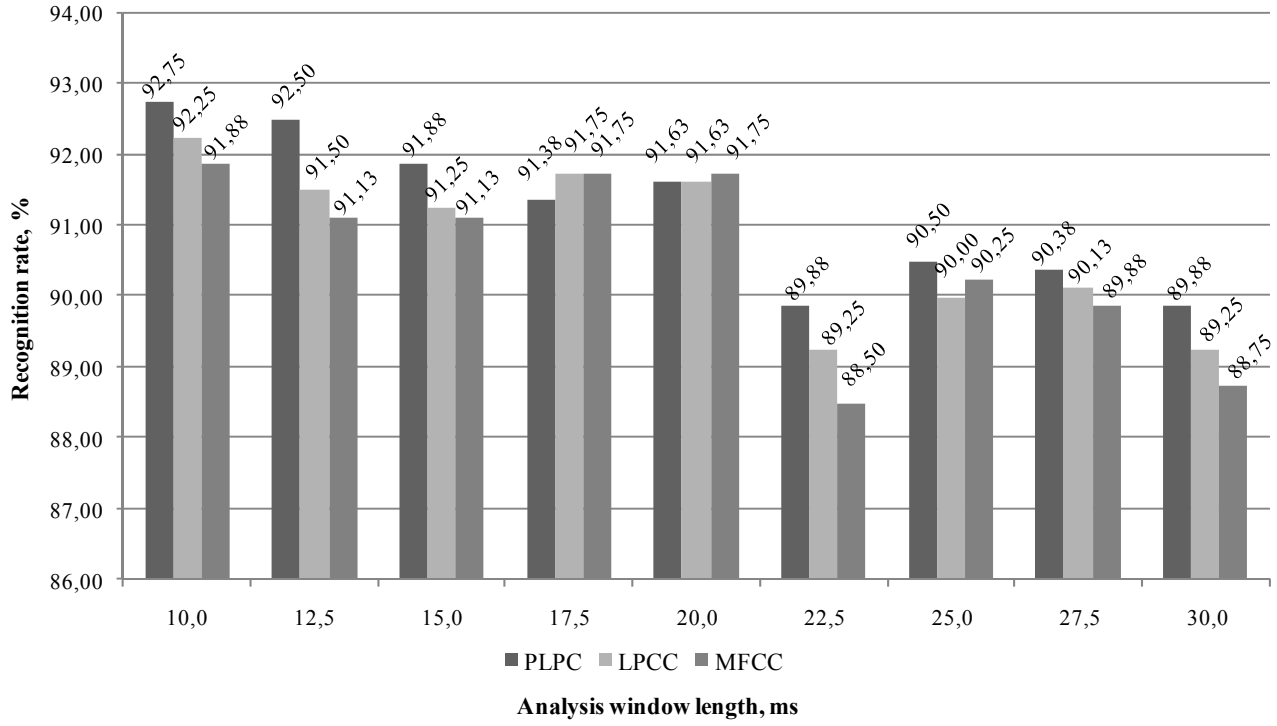

Fig. 2. Recognition rate dependence on analysis window length.

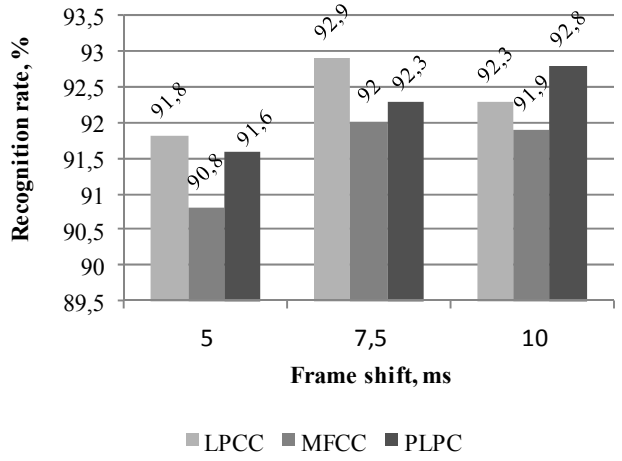

Fig. 3. Recognition rate dependence on analysis window shift.

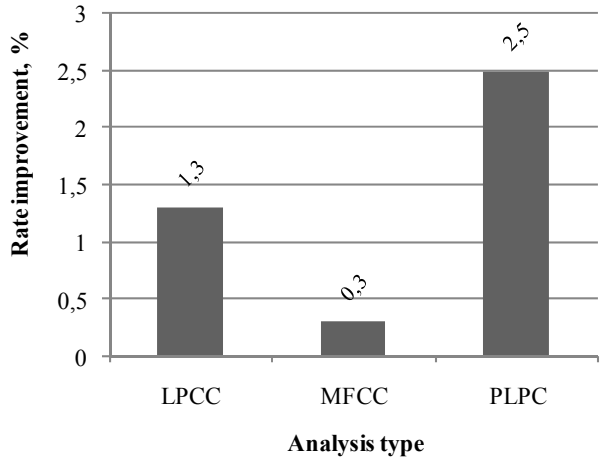

Fig. 4. Speech recognition improvement through analysis window parameters for different speech analysis techniques.

TABLE I

FRAME SHIFT VALUES GIVING HIGHEST RECOGNITION RATE

Analysis type	Minimal frame size		Maximal frame size	
	Frame size, ms	Frame shift, %	Frame size, ms	Frame shift, %
PLPC	10	100	15	83
LPCC	10	75	20	100
MFCC	10	75	20	100

The lowest recognition rates were obtained for the window length of 22.5–30 ms (27.5–30 ms for female speakers and 22.5 ms for male speakers particularly).

The second experiment was intended for frame shift selection. During this experiment we used the window length of 10 ms and varied frame shift from 5 ms (half-overlapped frames) to 10 ms (without overlapping). The recognition rate results for various frame shift values are given in Fig. 3.

In this case the highest recognition rate was obtained using 7.5 ms frame shift for MFCC and LPCC analyses. The highest recognition rate for PLPC analysis was obtained using 10 ms frame shift. So the frame shift equal to 75-100 % of window length allowed us to obtain the highest rate. The overall improvement of recognition rate caused by frame shift variation did not exceed 1 %.

For further study of frame shift we investigated recognition rate using maximal acceptable frame size values: 15 ms for PLP analysis, 20 ms for LPCC and MFCC analyses. Table I gives combined results of frame shift values giving the highest recognition rate for different frame size values (the frame shift values are expressed in terms of ratio to frame size). Here we can see the highest recognition rate obtained when using frame shift size equals to 75–100 % of the frame size. If the calculation speed is preferred the frame shift should be set to 100 % of the frame size.

Fig. 4 gives the overall improvement for all analysis techniques gained by changes of window length and frame shift size.

As we can see the accuracy of PLPC has increased by 2.5 % and this was the highest improvement (in comparison with the case of 20 ms window length and 10 ms frame shift). The MFCC analysis was most robust to change of analysis parameters – the recognition rate improvement was only 0.3 %.

In our experiments we registered different recognition rates for different analysis types and speakers, that is the increase/decrease of frame size and shift had diverse results for particular speaker. This implies the speaker-dependent optimal values of frame size and shift. Thus future research ought to be directed to adapt analysis parameters to speaker characteristics.

V. DISCUSSION AND CONCLUSIONS

The recognition rate dependence on analysis frame size and shift was investigated. The experimental results can be concluded:

- Shorter analysis frame can improve the average speech recognition rate. Twice shortened analysis frame (from 20 to 10 ms) gave us maximal recognition rate improvement by 1.2 %. The lowest average recognition rate was obtained for the window length of 22.5–30 ms.
- Longer frame shift accelerates speech analysis process, reduces the amount of extracted features and can improve speech recognition rate. The biggest recognition rate improvement was achieved for frame shift varying from 75 to 100 % of the frame size.
- The highest improvement of speech recognition rate (by 2.5 %) was achieved through variation of analysis parameters for PLP cepstral analysis. The MFCC analysis was the most robust for the analysis parameters variation.
- Considering the average recognition results for female and male speakers separately, adaptive analysis parameter values should be investigated for future improvement.

REFERENCES

[1] Z. Jiang, H. Huang, S. Yang, S. Lu, and Z. Hao, "Acoustic Feature Comparison of MFCC and CZT-Based Cepstrum for Speech Recognition," in Proceedings of 5th International Conference on Natural Computation, 2009, pp. 55–59.

[2] L. Deng, J. Wu, J. Droppo, and A. Acero, "Analysis and comparison of two speech feature extraction/compensation algorithms," IEEE Signal Processing Letters, vol. 12, no. 6, pp. 477–480, Jun. 2005.

[3] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[4] J. Pelecanos, S. Slomka, and S. Sridharan, "Enhancing automatic speaker identification using phoneme clustering and frame based parameter and frame size selection," in Proceedings of the 5th International Symposium on Signal Processing and its Applications ISSPA99 (IEEE Cat. No.99EX359), vol. 2, pp. 633–636.

[5] K. Paliwal and K. Wojcicki, "Effect of Analysis Window Duration on Speech Intelligibility," IEEE Signal Processing Letters, vol. 15, pp. 785–788, 2008.

[6] L. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition, 1st ed. Prentice Hall, 1993, p. 496.

[7] M. Goyani, N. Dave, and N. M. Patel, "Performance Analysis of Lip Synchronization Using LPC, MFCC and PLP Speech Parameters," in Proceedings of International Conference on Computational Intelligence and Communication Networks, 2010, pp. 582–587.

[8] M. Suzuki, T. Yoshioka, S. Watanabe, N. Minematsu, and K. Hirose, "MFCC enhancement using joint corrupted and noise feature space for highly non-stationary noise environments," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 4109–4112.

[9] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "Robust Speech Recognition Using a Cepstral Minimum-Mean-Square-Error-Motivated Noise Suppressor," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 5, pp. 1061–1070, Jul. 2008.

[10] T. Kinnunen, R. Saeidi, F. Sedlak, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, "Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 7, pp. 1990–2001, Sep. 2012.

[11] O. Gauci, C. J. Debono, and P. Micallef, "A nonlinear feature extraction method for phoneme recognition," in Proceedings of MELECON 2008 - The 14th IEEE Mediterranean Electrotechnical Conference, 2008, pp. 811–815.

[12] C. Lee, D. Hyun, E. Choi, J. Go, and C. Lee, "Optimizing feature extraction for speech recognition," IEEE Transactions on Speech and Audio Processing, vol. 11, no. 1, pp. 80–87, Jan. 2003.

[13] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," The Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738–1752, 1990.

[14] J. Makhoul, "Linear prediction: A tutorial review," Proceedings of the IEEE, vol. 63, no. 4, pp. 561–580, 1975.

[15] O. Rioul and M. Vetterli, "Wavelets and signal processing," IEEE Signal Processing Magazine, vol. 8, no. 4, pp. 14–38, Oct. 1991.

[16] M. Cutajar, E. Gatt, I. Grech, O. Casha, and J. Micallef, "Comparative study of automatic speech recognition techniques," IET Signal Processing, vol. 7, no. 1, pp. 25–46, Feb. 2013.

[17] A. Salomon, C. Y. Espy-Wilson, and O. Deshmukh, "Detection of speech landmarks: Use of temporal information," The Journal of the Acoustical Society of America, vol. 115, no. 3, pp. 1296–1305, 2004.

[18] U. H. Yapanel and J. H. L. Hansen, "A New perspective on Feature Extraction for Robust In-Vehicle Speech Recognition," in ISCA Proceedings: Eurospeech2003, 2003, pp. 1281–1284.

[19] C. Kim and R. M. Stern, "Power function-based power distribution normalization algorithm for robust speech recognition," in Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding, 2009, pp. 188–193.

[20] S. Kim, T. Eriksson, H.-G. Kang, and D. H. Youn, "A pitch synchronous feature extraction method for speaker recognition," in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. I–405–8.

[21] I. Ding, "Enhancement of speech recognition using a variable-length frame overlapping method," in Proceedings of International Symposium on Computer, Communication, Control and Automation (3CA), 2010, pp. 375–377.

[22] Q. Zhu and A. Abeer, "On the use of variable frame rate analysis in speech recognition," in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), vol. 3, pp. 1783–1786.

[23] B. Zhu and E. Micheli-Tzanakou, "Nonstationary speech analysis using neural prediction," IEEE Engineering in Medicine and Biology Magazine, vol. 19, no. 1, pp. 102–105, 2000.

[24] A. Lipeika, J. Lipeikiene, and L. Telksnys, "Development of Isolated Word Speech Recognition System," Informatica, vol. 13, no. 1, pp. 37–46, 2002.

[25] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, US. Prentice Hall, 1978, p. 962.

[26] K. K. Paliwal, J. G. Lyons, and K. K. Wojcicki, "Preference for 20-40 ms window duration in speech analysis," in Proceedings of 4th International Conference on Signal Processing and Communication Systems, 2010, pp. 1–4.

[27] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 25, no. 1, pp. 24–33, Feb. 1977.

[28] W.-G. Gong, L.-P. Yang, and D. Chen, "Pitch Synchronous Based Feature Extraction for Noise-Robust Speaker Verification," in Proceedings of Congress on Image and Signal Processing, 2008, pp. 295–298.

[29] G. L. Sarada, T. Nagarajan, and H. A. Murthy, "Multiple frame size and multiple frame rate feature extraction for speech recognition," in Proceedings of International Conference on Signal Processing and Communications, SPCOM '04., pp. 592–595.

[30] R. D. Zilca, B. Kingsbury, J. Navratil, and G. N. Ramaswamy, "Pseudo Pitch Synchronous Analysis of Speech With Applications to Speaker Recognition," IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 2, pp. 467–478, Mar. 2006.

[31] Z.-H. Tan and B. Lindberg, "Low-Complexity Variable Frame Rate Analysis for Speech Recognition and Voice Activity Detection," IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 5, pp. 798–807, Oct. 2010.

[32] C.-S. Jung, M. Y. Kim, and H.-G. Kang, "Selecting Feature Frames for Automatic Speaker Recognition Using Mutual Information," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 6, pp. 1332–1340, Aug. 2010.

Deividas Eringis received B.S., and M.S. degrees in electrical and electronic engineering from Vilnius Gediminas Technical University, Vilnius, Lithuania in 2008 and 2010, respectively. From 2011 he started to work in Vilnius University Institute of Mathematics and Informatics, to pursue research in speech recognition. He is currently pursuing the Ph.D. degree in informatics engineering. His current technical interests include speech signal processing, speech and pattern recognition. Address: Vilnius University, Institute of Mathematics and Informatics, Recognition Process Department, Lithuania, Vilnius LT-08663, Akademijos str., 4. E-mail: deividas.eringis@mii.vu.lt

Gintautas Tamulevičius received the B.S., M.S. in electronic engineering and Ph.D. degrees in informatics engineering from Vilnius Gediminas Technical University, Vilnius, Lithuania in 2001, 2003, and 2008 respectively.

He is Assistant Professor in the Department of Electronic Systems, Vilnius Gediminas Technical University, Vilnius, Lithuania. From 2001 to 2003, he worked at Vilnius University Institute of Mathematics and Informatics as a member of technical staff. From 2007 he is a Researcher in Vilnius University Institute of Mathematics and Informatics. His research interests include digital signal processing, speech recognition, pattern recognition. Address: Vilnius University, Institute of Mathematics and Informatics, Recognition Process Department, Lithuania, Vilnius LT-08663, Akademijos str., 4. E-mail: gintautas.tamulevicius@mii.vu.lt