ITC 3/48

Journal of Information Technology and Control
Vol. 48 / No. 3 / 2019
pp. 446-453
DOI 10.5755/j01.itc.48.3.21248

Speaker Discrimination Using Long-Term Spectrum of Speech

Received 2018/07/16 | Accepted after revision 2019/07/29

http://dx.doi.org/10.5755/j01.itc.48.3.21248

Speaker Discrimination Using Long-Term Spectrum of Speech

Milan Sigmund

Brno University of Technology, Institute of Radio Electronics, Technicka 12, CZ-61600, Brno, Czech Republic

Corresponding author: sigmund@feec.vutbr.cz

In this article, a specific long-term speech spectrum was investigated with respect to its use for speaker recognition. The long-term effect was satisfied by averaging short-term autocorrelation coefficients over the whole utterance. The long-term spectrum was calculated by means of second-order linear prediction using the average autocorrelation coefficients. First, speaker discriminability of 32 individual parameters was evaluated by combining spectral energy and spectral slope in eight different frequency bands covering the range 0–4 kHz (seven narrow nonoverlapping subbands and one band spanning over the full range). Then, four subbands with the most discriminative capability were selected for speaker recognition. These subbands involve the frequencies of 0–1.2 kHz in total. In the main experiments, text-independent speaker recognition based on relative Euclidean distance was performed in each single subband as well as in multiple 2 to 4 subbands applying two types of speech data, complete continuous speech and voiced part of the same speech. The voiced speech seems to be generally more effective for speaker recognition using the long-term speech spectrum. The best recognition rates, i.e. 91.7% on complete speech and 100% on voiced speech, were achieved in optimal paired subbands. The long-term speech spectrum can complement the traditional voice features.

KEYWORDS: Speech signal, Long-term spectrum, Efficient features, Speaker discrimination, Evaluation.

1. Introduction

One of the issues in speech signal processing as well as in biometric data mining is the investigation "How is the person's individuality reflected in voice?" There is no standard set of speech signal features commonly adopted for speaker recognition. Generally, automatic speaker recognition should be based on features which express both the speaking style and the anatomical structure of the speaker's vocal apparatus

[14]. A brief introduction into speaker recognition can be found in [22, 25]. A more comprehensive overview focused on various aspects of speaker recognition is given in [2]. An overview of technologies dealing with robustness related issues is provided in [30]. A practical introduction into phonetics and phonology is presented in the book [6]. Useful information for voice specialists contains the comprehensive textbook [20].

The long-term spectrum provides information on spectral energy distribution of a speech signal during a relatively long utterance. Such spectral characteristics make possible to reflect the anatomy and function of a vocal tract independently of current spoken phonemes. In paper [3], Byrne with his colleagues investigated 12 languages (namely English, French, German, Russian, Swedish, Danish, Welsh, Japanese, Cantonese, Mandarin, Singhalese, and Vietnamese) and compared their spectral characteristics. Although some statistically significant differences have been found among them, researchers concluded that all spectra are similar enough and they suggested a “universal” average spectrum of speech across all investigated languages. This normative spectrum can be used for many clinical objectives such as prescription and evaluation of hearing aid.

Some authors used long-term spectra for specific speech investigation. The study in [24] is focused on differentiating synthetic speech signals generated by vocoders from natural speech. Long-term-speech spectra of two different speaking styles, i.e. reading text and spontaneous speech, are compared in [8]. Differences by gender in long-term speech spectra of Turkish were observed and analyzed in [27]. The analysis in [10] presents gender differences of adult Iranian speakers. Opera singers performing singing, stage speech and conversational speech are analysed in [5]. Most of the studies concerning long-term spectra are based on processing the whole spectra obtained by the FFT approach as an averaged sequence of short-term speech spectra. The aim of this article is to investigate the discrimination power of the long-term spectrum for speaker recognition in the case where the spectrum is estimated using a low-order linear prediction.

The rest of this article is organized as follows. In Section 2, the applied algorithm for estimating the long-term spectrum is defined. Experimental setups are reported in Section 3 which is divided into four subsections. The first subsection describes speech materials used in our experiments. The second subsection defines individual spectral parameters used in the experiments. The third subsection explains the evaluation of all proposed parameters by means of discrimination power. The fourth subsection presents speaker recognition accuracy. Finally, Section 4 briefly concludes the article.

2. Estimation of the Long-Term Spectrum

We applied a well-known linear prediction approach which is often used for estimating short-term spectrum of phonemes in speech recognition. The standard approach was modified to directly estimate the long-term spectrum which is independent of linguistic context. Advantages of the modified method lie in its ability to directly provide a smoothed spectral envelope of long speech and its relative high speed of computation. The whole theory of linear prediction is presented in many books; for instance in [17]. In this section, we briefly describe the algorithm as it was applied in our experiments.

First, the speech signal was segmented into short frames of 20 ms using a rectangle window and in each frame the three lowest autocorrelation coefficients were computed. For the j-th frame of the speech signal $\{s(n)\}$, there are autocorrelation coefficients

$$R(j,k)=\sum_{n=1}^{N-k}s\left[(j-1)N+n\right]s\left[(j-1)N+n+k\right], \tag{1}$$

where k=0, 1, 2 stands for lag index corresponding to the time shift (given in samples) and N is the number of signal samples in each frame. The long-term averaging was made by means of short-term autocorrelation coefficients $R(j,k)$ averaged over the whole utterance consisting of J speech frames

$$\overline{R}(k)=\frac{1}{J}\sum_{j=1}^{J}R(j,k). \tag{2}$$

Then, two average predictive coefficients were computed from the coefficients $\overline{R}(k)$ as follows

$$\bar{a}_1 = \frac{\bar{R}(1)\bar{R}(0) - \bar{R}(1)\bar{R}(2)}{\bar{R}(0)^2 - \bar{R}(1)^2}, \quad (3a)$$

$$\bar{a}_2 = \frac{\bar{R}(2)\bar{R}(0) - \bar{R}(1)\bar{R}(1)}{\bar{R}(0)^2 - \bar{R}(1)^2}. \quad (3b)$$

Finally, the long-term magnitude spectrum was estimated by

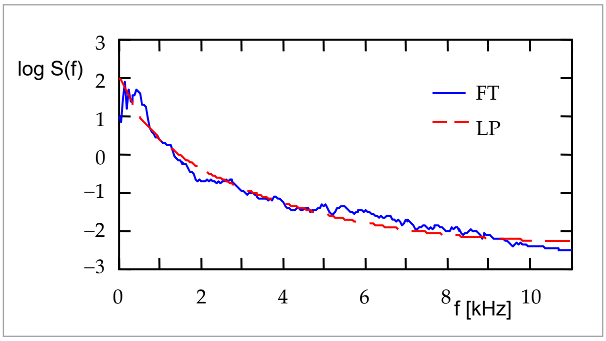
$$S(f) = \left|1 - \bar{a}_1 \exp(-i2\pi f / f_s) - \bar{a}_2 \exp(-i4\pi f / f_s)\right|^{-2}, \quad (4)$$

where f_s denotes the sampling rate of speech signal and i stands for imaginary unit. In general, the spectrum can be computed for a frequency with a sweep from f=0 Hz up to the Nyquist frequency f_s/2. The number of predictive coefficients used to compute the speech spectrum (i.e., the order of linear prediction) can effectively control the degree of spectral smoothness. The full algorithm for estimating the long-term spectrum is computationally very efficient due to the determination of long-term characteristics at the low feature level (using autocorrelation coefficients).

An important factor for practical applications seems to be the needed length of the utterance in order to represent the speakers independently of the spoken text. Experimental results show that speech of approximately 100 seconds (i.e., 5000 voiced and unvoiced speech frames) satisfies statistical reliability. This corresponds to the steadying of some relevant individual features characterizing speakers such as mean value of voice fundamental frequency [21]. Figure 1 illustrates a typical long-term spectrum estimated from the same speech signal using two different approaches. The spectrum computed by the above defined algorithm based on linear prediction is graphically compared here with the long-term spectrum obtained by Fourier transform. The spectral values are displayed logarithmically in order to see small differences in the whole frequency range from 0 to 11 kHz. The second order of linear prediction satisfies the best fitting to the long-term FT-spectrum. Moreover, estimation of the LP-spectrum of second order is computationally very simple.

Figure 1

Comparison of the long-term spectra obtained by Fourier transform (FT) and by linear prediction (LP)



3. Experimental Results

The purpose of carrying out experiments is to find out which of the long-term spectral parameters are best suited for speaker discrimination. The attention is focused on spectral energy and spectral slope in subband processing. If any subband is corrupted with strong noise, the other subbands may be explored.

3.1. Speech Material

All experiments were conducted on Czech speech signals. In the initial measurements, the spectra from 18 male speakers (Czech natives) were investigated. The speakers were asked to read an identical text which is phonetically balanced. In that case, a speech duration of approximately 1.5 minutes was regarded as sufficient enough to get long-term spectra independent of the text. All speakers were instructed to read the text in their natural voice including normal reading tempo and habitual loudness. No speakers indicated pronunciation problems or illness, such as symptoms of cold or acute respiratory infection. The speech was recorded in a quiet environment at 22 kHz and 16 bits per sample using a standard personal computer equipped with an internal sound card and external microphone with very linear frequency response (Behringer ECM 8000). For the experimentation with long-term characteristics, the stored speech signal was resampled at 8 kHz.

3.2. Analysed Parameters

The long-term spectrum obtained was divided into 7 adjacent subbands without overlapping. The subbands were nonlinearly spaced on the frequency axis (no frequency warping) according to the curvature of the long-term spectrum averaged across all speakers (see Figure 2) as follows: 0–200 Hz, 200–500 Hz, 500–800 Hz, 800–1200 Hz, 1.2–2 kHz, 2–3 kHz, and 3–4 kHz. Each subband was considered independently for further computations. Generally, the spectral variability between speakers seems to be most significant in the lower part of the long-term spectrum.

Figure 2

Long-term spectrum averaged across all speakers in the frequency range 0–4 kHz

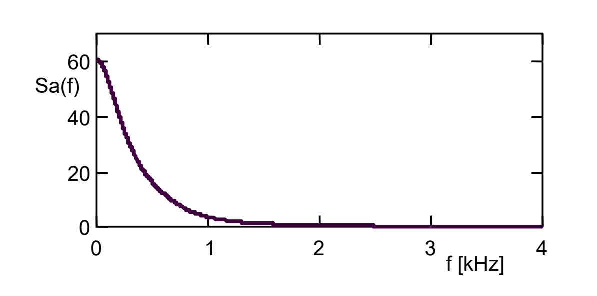
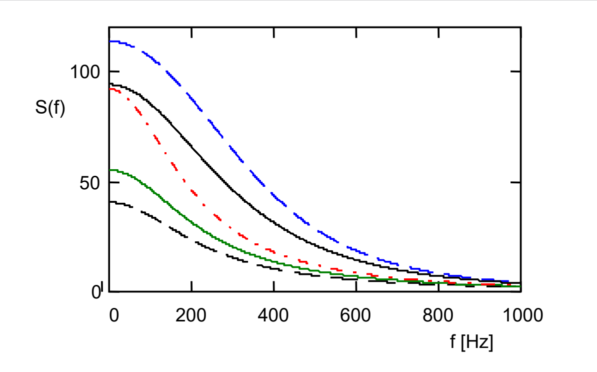


Figure 3 shows, in detail, individual long-term spectra (non-logarithmic) for five speakers at low frequencies up to 1000 Hz. All spectra (also those not shown in Figure 3) decrease monotonously.

Figure 3

Examples of long-term spectra from five speakers in the low frequency range 0–1 kHz



Both non-logarithmic and logarithmic spectra were considered in the primary investigation. In each subband as well as in the full frequency band 0–4 kHz, spectral energy and spectral slope were calculated. Thus, altogether 32 parameters were analysed (16 parameters in non-logarithmic spectrum and the same 16 ones in logarithmic spectrum).

The spectral energy E related to a bandwidth with cutoff frequencies $f_{\min}$ and $f_{\max}$ was calculated using modified Parseval's theorem [15]

$$E = \int_{f_{\min}}^{f_{\max}} S^2(f)\, df. \qquad (5)$$

The second parameter, spectral slope K, represents the slope of a straight line that best approximates the long-term spectrum in the respective subband (see dashed line in Figure 4). The optimal approximation was found using the mean-square error criterion [26].

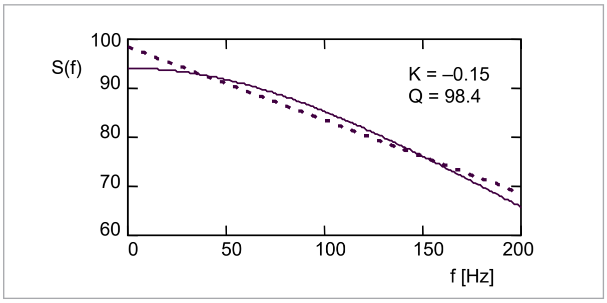
For calculating K, the usual equation of a straight line in x-y-plane written as

$$y = Q + K\,x \qquad (6)$$

was used, where Q is the point where the line crosses the vertical axis and K is a coefficient of direction, i.e. the sought slope. When $K<0$, the line has a descending trend, while $K>0$ means an ascending trend.

Figure 4

Illustration of spectral slope determination in the subband 0–200 Hz



3.3. Evaluation of Used Parameters

In order to select the best subband as well as the best individual parameter for speaker discrimination, the measure $V(x)$ based on variability was applied for evaluation

$$V(x) = \frac{\sum_i |x_i - \mu(x)|}{|\mu(x)|}, \qquad (7)$$

where x represents each of the 32 parameters described in the previous section, $\mu(x)$ stands for mean of the parameter x and i is index for speakers. When the values of x_i differ significantly between speakers, criterion $V(x)$ gets higher. The discrimination power of all individual parameters was calculated across all speakers and then the values of $V(x)$ for energy and slope were averaged in each subband. Table 1 shows frequency bands ranked in terms of average $V(x)$.

Table 1

Ranked subbands of long-term spectrum

Rank	Band [kHz]	Criterion V(x)		
		Energy	Slope	Average
1	0.2–0.5	6.286	**5.308**	5.797
2	0.5–0.8	**7.154**	4.193	5.674
3	0–0.2	6.227	4.412	5.320
4	0.8–1.2	5.988	4.119	5.053
5	0–4	5.816	3.153	4.485
6	3–4	4.357	1.103	2.730
7	1.2–2	2.630	1.845	2.237
8	2–3	3.352	0.655	2.003
mean		5.23	3.19	4.16

Table 2 shows the discriminative ranking of frequency bands when the logarithmic spectrum was used.

Table 2

Ranked subbands of logarithmic long-term spectrum

Rank	Band [kHz]	Criterion V(x)		
		Energy	Slope	Average
1	0–0.2	3.007	**4.703**	3.855
2	0.2–0.5	2.920	3.928	3.424
3	0.8–1.2	**4.323**	1.876	3.099
4	0.5–0.8	4.030	1.591	2.811
5	1.2–2	3.580	1.564	2.572
6	2–3	3.222	1.145	2.183
7	0–4	2.970	1.245	2.108
8	3–4	3.118	0.955	2.036
mean		3.40	2.13	2.76

In both tables, the highest values of each parameter representing the best discrimination power are highlighted in bold typeface. Accordingly, the overall best individual parameter was energy in the subband 0.5–0.8 kHz of the non-logarithmic spectrum which achieved a score of $V(x)$=7.154. Generally, comparing the overall averages of $V(x)$ between non-logarithmic and logarithmic spectrum, i.e. 4.16 versus 2.76 (see the lower right corners in Tables 1 and 2), the non-logarithmic spectrum offers better discrimination for our purpose. In both kinds of spectrum, lower subbands are better suited for speaker discrimination than higher subbands, as well as the full band of 0–4 kHz. The good efficiency of the low-frequency part of the spectrum in the range of 0–1.2 kHz is related to some significant frequency features characterizing the voice [1, 23] such as fundamental frequency in the range 80–160 Hz for male speakers or 120–300 Hz for female speakers as well as low formants of vowels in the range from 400 Hz above.

3.4. Speaker Recognition

In the final experiments, the above described parameters were applied for text-independent speaker recognition within a group of 12 male Czech native speakers. With respect to the discrimination power expressed in Tables 1 and 2, the most efficient parameters were examined for speaker recognition. Thus, the applied parameters are spectral energies and spectral slopes in four subbands covering the frequency range 0–1.2 kHz of the non-logarithmic spectrum. The speakers spoke different texts in the training and testing phases (each text of approximately 1.5 minutes). The recognition process was based on squared Euclidean distance which quantifies the similarity between an unknown speaker and each known speaker in the system database. The distance of tested (i.e. unknown) speaker to each known speaker was calculated as

$$D = \sum_{band} \sum_{parameter} \frac{(x^{\mathrm{R}} - x)^2}{(x^{\mathrm{R}})^2}, \tag{8}$$

where x represents the parameters energy E and slope K, respectively. The superscript R indicates the reference parameters of known speakers. The reference values were obtained and stored in system database during training. Due to the very unbalanced numerical values of both parameter types (e.g., in the lowest subband $E>10^7$ and $K<-1$), each individual difference $(x^R - x)^2$ is scaled by the denominator $(x^R)^2$ before summing in Equation (8). In the case of the simplest recognition based on one parameter extracted from one subband, Equation (8) can be reduced to the form

$$D = (x^R - x)^2 . \tag{9}$$

During the experiments, many speaker recognition tests were performed, combining both groups of factors: band parameters and different subbands. The recognition accuracy was estimated in all tests as the ratio

$$accuracy = \frac{correctly\ \ recognised\ \ speakers}{total\ \ number\ \ of\ \ speakers} . \tag{10}$$

When using energy and/or slope extracted from only one individual subband, the recognition accuracy seems to be insufficient. In this case, the best recognition rate of 75% was achieved in the subband 0–200 Hz. Thus, there is no narrow single subband applicable for practical tasks in speaker recognition. However, the results will improve significantly when using multiple subbands. The recognition scores achieved for various combinations of 2 to 4 subbands and one or two parameters are summarized in Table 3. As can be seen, the best recognition rate of 91.7% was achieved in seven different combinations of subbands. All these combinations contain the lowest subband 0–200 Hz. The 91.7%-result was achieved with both parameters (energy and slope), but also with one parameter (energy or slope).

Further, the tests as presented in Table 3 were performed newly with voiced speech instead of complete speech. In the case of continuous speech processing, the voiced speech (i.e., voiced phonemes only) is the most common alternative to the complete speech (i.e., all phonemes, voiced and unvoiced). The ratio of voiced and unvoiced speech frames depends on speaking style, speaker environment as well as national language. In the used speech signals, the voiced frames cover on average 58% of all frames of the complete speech.

Table 3

Speaker recognition rates (in percent) based on continuous speech using multiple subbands

Range [kHz]	Number of subbands	Energy	Slope	Energy and Slope
0–0.5	2	75.0	91.7	91.7
0–0.8	3	83.3	91.7	91.7
0.2–0.8	2	58.3	58.3	58.3
0.2–1.2	3	58.3	66.7	58.3
0.5–1.2	2	41.7	33.3	41.7
0–0.2 & 0.5–0.8	2	91.7	91.7	91.7
0–0.2 & 0.8–1.2	2	91.7	91.7	91.7
0–0.2 & 0.5–1.2	3	83.3	91.7	91.7
0.2–0.5 & 0.8–1.2	2	58.3	66.7	66.7
0–0.5 & 0.8–1.2	3	83.3	91.7	91.7
0–1.2	4	83.3	91.7	91.7

Table 4

Speaker recognition rates (in percent) based on voiced speech using multiple subbands

Range [kHz]	Number of subbands	Energy	Slope	Energy and Slope
0–0.5	2	91.7	91.7	100
0–0.8	3	91.7	100	91.7
0.2–0.8	2	75.0	83.3	75.0
0.2–1.2	3	75.0	83.3	75.0
0.5–1.2	2	58.3	66.7	66.7
0–0.2 & 0.5–0.8	2	91.7	91.7	100
0–0.2 & 0.8–1.2	2	91.7	91.7	100
0–0.2 & 0.5–1.2	3	91.7	100	91.7
0.2–0.5 & 0.8–1.2	2	75.0	91.7	83.3
0–0.5 & 0.8–1.2	3	91.7	100	91.7
0–1.2	4	91.7	91.7	91.7

The long-term spectrum of voiced speech was estimated by the same approach as described in Section 2, but in Equation (2) autocorrelation coefficients $R(j,k)$ extracted solely from voiced frames of the speech signal were taken into account. The obtained results in Table 4 indicate that the voiced speech is better suited for speaker discrimination than the complete speech according to both mean score and the best individual score. In both speech variants, the best recognition rates, i.e. 91.7% on complete speech and 100% on voiced speech, were achieved in the subbands grouped as follows: 0–200 Hz with 200–500 Hz, 0–200 Hz with 500–800 Hz, and 0–200 Hz with 800–1200 Hz. In some subbands of voiced speech, the slope alone provides better recognition accuracy than both parameters energy and slope together.

4. Conclusion and Future Work

The presented experiments were aimed at comparing differences in speaker specific long-term spectra and their utilization for speaker recognition. In a group of 12 speakers, the highest recognition accuracy of 100% was achieved using spectral energy and spectral slope of voiced speech in three different combinations of subbands. This gives the possibility to prioritize a subband not affected by noise. The results of the research may be generalized to a new finding that one efficient parameter (here the spectral slope) derived from a suitable subband of smoothed long-term spectrum is sufficient to successfully discriminate against speakers.

When recognizing speakers and having long utterances available, the long-term speech spectrum can complement the traditional short-term voice features such as pitch [13], mel-frequency cepstral coefficients [11], line spectral pair frequencies [19], etc. and so help to improve recognition systems.

In future work, the long-term speech spectrum and extracted spectral parameters will be tested for their robustness to various factors affecting speech, such as emotions [7], physical fatigue [12], vocal effort [9, 28], and others [4]. Furthermore, it will be useful to also investigate the influence of adverse acoustic conditions, especially non-stationary noises [29] due to their indirect influence on speech production. Contrary to speaker recognition, it may be interesting to investigate the applicability of long-term spectra for spectral normalization in the field of speaker de-identification [16]. Recently, there has been a growing need for de-identification of multimedia data [18], in order to ensure their anonymity with respect to privacy protection in the European Union.

Acknowledgement

Research described in this article was financed by the Czech Ministry of Education within the frame of the National Sustainability Program under grant LO1401. For research, infrastructure of the SIX Center was used.

References

1. Baken, R., Orlikoff, R. Clinical Measurement of Speech and Voice. Singular Publishing, San Diego, 2000.
2. Beigi, H. Fundamentals of Speaker Recognition. Springer, New York, 2011. https://doi.org/10.1007/978-0-387-77592-0
3. Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Kotby, M. N., Nasser, N. H. A., El Kholy, W. A. H., Nakanishi, Y., Oyer, H., Powell, R., Stepehens, D., Meredith, R., Sirimanna, T., Tavartkiladze, G., Frolenkov, G. I., Westerman, S., Ludvigsen, C. An International Comparison of Long-Term Average Speech Spectra. Journal of the Acoustical Society of America, 1994, 96(4), 2108-2120. https://doi.org/10.1121/1.410152
4. Čeidaite, G., Telksnys L. Analysis of Factors Influencing Accuracy of Speech Recognition. Elektronika ir Elektrotechnika, 2010, 105(9), 69-72. https://doi.org/10.5755/j01.eee.105.9.9180
5. Dong, L., Kong, J., Sundberg, J. Long-Term-Average Spectrum Characteristics of Kunqu Opera Singers' Speaking, Singing and Stage Speech. Logopedics Phoniatrics Vocology, 2014, 39(2), 72-80. https://doi.org/10.3109/14015439.2013.841752
6. Fromkin, V., Rodman, R., Hyams, N. An Introduction to Language (11th edition). Cengage Learning, Boston, 2019.
7. Gangamohan, P., Kadiri, S. R., Yegnanarayana, B. Analysis of Emotional Speech-A Review. In: Esposito, A., Jain, L. (Eds.), Toward Robotic Socially Believable Behaving Systems, Springer, Cham, 2016, 205-238. https://doi.org/10.1007/978-3-319-31056-5_11
8. Lee, K., Jin, I. K. Comparison of Long-Term Average Speech Spectra in Reading Context and Spontaneous Speech. Clinical Archives of Communication Disorders, 2017, 2(1), 85-89. https://doi.org/10.21849/cacd.2016.00115
9. López, A. R., Saeidi, R., Juvela, L., Alku, P. Normal-To-Shouted Speech Spectral Mapping for Speaker Recognition under Vocal Effort Mismatch. Proceedings of IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP), New Orleans, 2017, 4940-4944. https://doi.org/10.1109/ICASSP.2017.7953096

10. Moradi, N., Maroufi, N., Bijankhan, M., Nik, T. H., Salavati, M., Jalayer, T., Latifi, S. M., Soltani, M. Long-Term Average Spectra of Adult Iranian Speakers' Voice. Journal of Voice, 2014, 28(3), 305-310. https://doi.org/10.1016/j.jvoice.2013.09.002

11. Muttaqi, T., Mousavinezhad, S. H., Mahamud, S. User Identification System Using Biometrics Speaker Recognition by MFCC and DTW Along with Signal Processing Package. Proceedings of IEEE International Conference on Electro/Information Technology (EIT), Rochester, US, 2018, 79-83. https://doi.org/10.1109/EIT.2018.8500256

12. Nascimento, C. L., Constantini, A. C., Mourão, L. F. Vocal Effects in Military Students Submitted to an Intense Recruit Training: A Pilot Study. Journal of Voice, 2016, 30(1), 61-69. https://doi.org/10.1016/j.jvoice.2015.03.005

13. Nasr, M. A., Abd-Elnaby, M., El-Fishawy, A. S., El-Rabaie, S., El-Samie, F. E. A. Speaker Identification Based on Normalized Pitch Frequency and Mel Frequency Cepstral Coefficients. International Journal of Speech Technology, 2018, 21(4), 941-951. https://doi.org/10.1007/s10772-018-9524-7

14. Orsag, F. Speaker Recognition in the Biometric Security Systems. Computing and Informatics, 2006, 25(5), 369-391.

15. Paul, C. R. Transmission Lines in Digital and Analog Electronic Systems. John Wiley & Sons, Hoboken, New Jersey, 2010. https://doi.org/10.1002/9780470651414

16. Pribil, J., Pribilova, A., Matousek, J. Evaluation of Speaker De-Identification Based on Voice Gender and Age Conversion. Journal of Electrical Engineering, 2018, 69(2), 138-147. https://doi.org/10.2478/jee-2018-0017

17. Rabiner, L. R., Schafer, R. W. Theory and Applications of Digital Speech Processing. Prentice Hall, London, 2011.

18. Ribaric, S., Ariyaeeinia, A., Pavesic, N. De-Identification for Privacy Protection in Multimedia Content: A Survey. Signal Processing: Image Communication, 2016, 47, 131-151. https://doi.org/10.1016/j.image.2016.05.020

19. Sahidullah, M., Chakroborty, S., Saha, G. On the Use of Perceptual Line Spectral Pairs Frequencies and Higher-Order Residual Moments for Speaker Identification. International Journal of Biometrics, 2010, 2(4), 358-378. https://doi.org/10.1504/IJBM.2010.035450

20. Sataloff, R. T. Professional Voice: The Science and Art of Clinical Care (4th edition). Plural Publishing, San Diego, 2017.

21. Sigmund, M. Statistical Analysis of Fundamental Frequency Based Features in Speech Under Stress. Information Technology and Control, 2013, 42(3), 286-291. https://doi.org/10.5755/j01.itc.42.3.3895

22. Singh, S. Forensic and Automatic Speaker Recognition System. International Journal of Electrical and Computer Engineering, 2018, 8(5), 2804-2811. https://doi.org/10.11591/ijece.v8i5.pp2804-2811

23. Stanek, M., Sigmund, M. Finding the Most Uniform Changes in Vowel Polygon Caused by Psychological Stress. Radioengineering, 2015, 24(2), 604-609. https://doi.org/10.13164/re.2015.0604

24. Tian, X., Du, S., Xiao, X., Xu, H., Chng, E., Li, H. Detecting Synthetic Speech Using Long Term Magnitude and Phase Information. Proceedings of IEEE China Summit and International Conference on Signal and Information Processing, Chengdu, 2015, 611-615. https://doi.org/10.1109/ChinaSIP.2015.7230476

25. Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., Wang, R. Speaker Identification Features Extraction Methods: A Systematic Review. Expert Systems with Applications, 2017, 90, 250-271. https://doi.org/10.1016/j.eswa.2017.08.015

26. Varatharajan, R., Manogaran, G., Priyan, M. K. A Big Data Classification Approach Using LDA with an Enhanced SVM Method for ECG Signals in Cloud Computing. Multimedia Tools and Applications, 2018, 77(8), 10195-10215. https://doi.org/10.1007/s11042-017-5318-1

27. Yüksel, M., Gündüz, B. Long Term Average Speech Spectra of Turkish. Logopedics Phoniatrics Vocology, 2018, 43(3), 101-105. https://doi.org/10.1080/14015439.2017.1377286

28. Zelinka, P., Sigmund, M. Automatic Vocal Effort Detection for Reliable Speech Recognition. Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Kittila, Finland, 2010, Article Number: 5589174, 349-354. https://doi.org/10.1109/MLSP.2010.5589174

29. Zelinka, P., Sigmund, M. Hierarchical Classification Tree Modeling of Nonstationary Noise for Robust Speech Recognition. Information Technology and Control, 2010, 39(3), 202-210. https://doi.org/10.5755/j01.itc.39.3.12376

30. Zheng, T. F., Li, L. Robustness-Related Issues in Speaker Recognition. Springer Nature, Singapore, 2017. https://link.springer.com/content/pdf/10.1007/ 978-981-10-3238-7.pdf