ALFA BK UNIVERSITY
FACULTY OF INFORMATION TECHNOLOGY
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
ALFATECH Journal

ISSN 1XX0-3XX1(Print), ISSN 1XXX-6XXX (Online)

Alfa BK Univerzitet

UDK: 621.391:681.84
004.7:711.417.4
DOI: 10.46793/AlfaTech1.1.05J
Original scientific paper

Application of Mel-Frequency Cepstral Coefficients in Automatic Speaker Recognition as Part of IoT Solutions for Security and Optimization in Smart Cities

Ivan JOKIĆ[1], Vlado DELIĆ[2], Zoran PERIĆ[3]

Abstract: This paper presents an implementation of automatic speaker recognition utilizing feature vectors composed of 21 mel-frequency cepstral coefficients (MFCCs) as part of an IoT-driven solution for enhancing security and optimization in smart cities. Experiments are conducted on the Solo portion of the CHAINS database, containing 33 unique sentences pronounced by each of 36 speakers. Results indicate that recognition accuracy varies with the training and testing datasets and improves with longer test recordings. A comparative analysis of MFCC calculation methods reveals that accuracy is generally higher when a sigmoidal square of amplitude characteristic is applied to frequency-selective ranges, rather than an exponential approach. Models are developed for each speaker's recordings, represented by a covariance matrix of feature vectors, and applying a sigmoid function to the model elements yields a 5% increase in recognition accuracy in most cases. These findings highlight the potential for MFCC-based speaker recognition as a scalable, data-driven IoT tool for security, public safety, and resource optimization in the context of smart cities.

Keywords: Automatic speaker recognition; Mel-frequency cepstral coefficients (MFCCs); Covariance matrix; Exponential; Sigmoidal.

1 INTRODUCTION

Feature determination is a basic step in automatic recognition. Recognitions of targeted information from speech, or from sound in general, can be realized by observing and analyzing spectrum of speech or sound. Mel-frequency cepstral coefficients (MFCCs) track the spectral envelope of speech or sound. These features are widely used in various fields of signal analysis applications: acoustic, medicine, industry [1]. Speech is a complex acoustic signal which contains information about: spoken textual content, identity of speaker, emotional state of speaker. Automatic speech recognition converts speech to text, in [2-4] MFCCs are used as features of speech. Aim of automatic speaker recognition is to recognize identity of speaker who speak. Different speakers have different voice color. Voices of different speakers differ in color. Sense of voice color is consequence of harmonic structure of spectrum of observed speech. In the field of number values, color of voice can be described as harmonic structure of spectrum. Information about harmonic structure of speech are contained in spectrum envelope of speech signal i.e. in appropriate MFCCs. Therefore, MFCCs as represents of spectral envelope and color of voice are also used in automatic speaker recognition [5-7]. Presence of emotion in speech additionaly colors the voice color of observed speech. Envelope of spectrum of observed speech i.e. appropriate MFCCs contain information about emotion present in speech [8-11]. Each sound in nature has appropriate spectral content. MFCCs as derived from spectral envelope are used in recognition of sound in nature [12-16].

MFCCs are short-term features of voice, usually of speech frames in duration around 25 ms in applications for speaker recognition. Next parts of procedure for automatic speaker recognition make compact representations i.e. models of short-term features. Models are long-term features of voice. They represent a larger sequence of speech frames, for example it can be pronunciation of one sentence in duration of around 2 seconds. It is of significance to have efficient short-term features. Their efficiency can be improved by adjusting the parameters observed during the calculation. In calculation of MFCCs one of parameters is energy observed in speech signal. MFCCs depend of energy spectrum of speech signal observed but also and of a manner how this spectrum is observed during calculation of MFCCs. Frequency selective filters are applied during calculation of MFCCs. Shape of square of amplitude characteristic of these filters has impact to calculated MFCCs and to achieved recognition accuracy [17]. These filters are used to model auditory critical bands, frequency range in which the listener does not distinguish a change in frequency between the two tones. Concept of auditory critical bands is related to the masking phenomena. Modeling of their amplitude characteristics can be done by nonlinear functions [18]. From the standpoint of listener, color of voice is constant property of speaker. From the standopint of automatic speaker recognizer, color of voice is calculated number or set of numbers. It is necessary that this number or set of numbers to be constant for one speaker. These values are also influenced by the textual content. To avoid this dependency, nonlinear modifications on used frequency selective bands, observed during MFCCs calculation, are applied.

In the rapidly evolving landscape of smart cities, solutions for enhanced security and efficient resource optimization are essential. Leveraging automatic speaker recognition within IoT frameworks offers a novel approach to addressing these needs by providing robust, real-time identification capabilities that enhance urban safety and operational efficiency.

In the continuation of the work, automatic speaker recognizer used is decsribed in section 2, speech database used and results of recognition accuracy are shown in section 3. Section 4 is devoted to conclusion remarks.

2 AUTOMATIC SPEAKER RECOGNIZER USED

Procedure for speaker recognition is based on use of short-term speech features. Experiments that will be described in this paper are conducted on the speech recordings whose frequency sampling is f_s=44100 Hz. Frames duration is 1024 samples, it is around 23,2 ms. They are mutually shifted by 368 samples or approximately by 8,3 ms. Speech frames are windowed by Hann window function. Feature vectors consist of 21 MFCCs.

MFCCs are calculated by equality:

$$c_n = \sum_{m=1}^{22} \log(E_m) \cdot \cos\left[\frac{\pi}{22} \cdot n \cdot \left(m - \frac{1}{2}\right)\right], \quad (1)$$

where n={1,2,...,21}, E_m is the energy in *m-th* frequency selective range. Here is used 22 frequency selective ranges, which are of 300 mel wide and mutually shifted by 150 mel. The boundaries of the first frequency selective range are 0 mel and 300 mel, boundaries of the second are 150 mel and 450 mel, so that boundaries of the twenty second range are (22-1)·150 mel = 3150 mel and 3150+300 mel = 3450 mel. MFCCs mirror the spectral content of sound signal in the space of MFCCs in accordance with human perception of spectral content. Energies contained in observed frequency selective ranges are parameters which have direct impact on values of MFCCs. Square of amplitude characteristic of frequency selective ranges has impact to energy E_m used for calculation of MFCCs. In [17] maximum of recognition accuracy is achieved for exponential type of this characteristic. Here recognition accuracy will be compared for exponential and sigmoidal type of this characteristic.

The exponential square of amplitude characteristic of frequency selective filters here used is:

$$A_m(k)^2 = \begin{cases} e^{(k-k_{c,m})\cdot 2}, & k_{1,m} \le k \le k_{c,m}, \\ e^{-(k-k_{c,m})\cdot 2}, & k_{c,m} < k \le k_{2,m}. \end{cases} \quad (2)$$

The sigmoidal square of amplitude characteristic of frequency selective filters here used is:

$$A_m(k)^2 = \begin{cases} sigm_1(k - k_{c,m}), & k_{1,m} \le k < k_{c,m}, \\ 1, & k = k_{c,m}, \\ sigm_1(k_{c,m} - k), & k_{c,m} < k \le k_{2,m} \end{cases} \quad (3)$$

Sigmoid function used is $sigm_1(x) = \frac{1}{1+e^{-a\cdot x}}$, parameter a is varied in two values: a=1 and a=0,5. Discrete frequencies: $k_{1,m}$, $k_{2,m}$ and $k_{c,m} = \frac{k_{1,m} + k_{2,m}}{2}$ are the lower, upper and central frequency of *m-th* frequency selective range respectively. Estimation of energy in observed frequency selective range is done by equality:

$$E_m = 2 \cdot \sum_{k=k_{1,m}}^{k_{2,m}} |X(k)|^2 \cdot A_m(k)^2, \quad (4)$$

$X(k)$ is discrete Fourier transform of observed speech frame. Short-term feature vectors of one recording of speaker are written into matrix of feature vectors X. For matrix X of n feature vectors appropriate model i.e. long-term feature of speaker is calculated as appropriate covariance matrix of matrix X,

$$\Sigma_{d\times d} = \frac{1}{n-1} \cdot \left(X_{d\times n} - \mu_{d\times 1}\right) \cdot \left(X_{d\times n} - \mu_{d\times 1}\right)^{\mathrm{T}}, \quad (5)$$

μ is vector of mean values of matrix X. Difference between two models, some i-th model Σ_i and observed reference model Σ_{ref}, is calculated by equality:

$$r(\Sigma_i, \Sigma_{ref}) = \frac{1}{d^2} \cdot \sum_{j=1}^{d} \sum_{k=1}^{d} \left|\Sigma_i(j,k) - \Sigma_{ref}(j,k)\right|, \quad (6)$$

where d is dimensionality of feature vector used, in experiments described in this paper d=21. Similar measure of difference between models is used in [19].

The speaker recognition procedure is adapted to work on closed set of speakers. In training for each of speech recordings is calculated appropriate matrix of feature vectors and covariance matrix. Models i.e. covariance matrices of the same speaker are named by the identity of that speaker. During testing for test speech recording is formed matrix of feature vectors and appropriate model i.e. covariance matrix Σ_{test}. Covariance matrix i.e. model of test speech recording is compared with all models determined in training. The identity of the most similar reference model is recognized speaker. If during training is formed M reference models, then test speech recording, *test*, have the identity of *i-th* reference model if $r(\Sigma_{test}, \Sigma_i) < r(\Sigma_{test}, \Sigma_j), \quad j \in \{1,2,...,M\} \setminus \{i\}$.

3 EXPERIMENT SETUP AND RESULTS

Experiments of speaker recognition are conducted on the Solo part of the CHAINS (CHAaracterizing INdividual Speakers) speech database [20]. The characteristic of this part of CHAINS database is: each of 36 speakers simply read a prepared text at a comfortable rate. Experiments are done on pronunciations of the same set of 33 individual sentences for each of speakers. Testing is organized in 12 tests. In each of 12 tests test set contains 3 recordings per speaker, rest of 30 recordings is used for training in that test. Recordings in CHAINS are recorded in WAV format, frequency sampling is 44100Hz and quantisation resolution is 16 bit/sample.

Table 1 Results for feature vector 21 MFCCs, exponential and sigmoidal square of amplitude of frequency selective ranges, minimum and maximum values of accuracy in each of columns are shaded

Test set	Exponential	Sigm. a=1,0	Sigm. a=0,5
s01,s03,s05	100/108 92,59%	98/108 90,74%	99/108 91,67%
s02,s04,s06	90/108	92/108	92/108

	83,33%	85,18%	85,18%
s07,s09,s11	93/108 86,11%	96/108 88,89%	97/108 89,81%
s08,s10,s12	104/108 96,3%	105/108 97,22%	105/108 97,22%
s13,s15,s17	103/108 95,37%	105/108 97,22%	104/108 96,3%
s14,s16,s18	105/108 97,22%	105/108 97,22	105/108 97,22%
s19,s21,s23	104/108 96,3%	105/108 97,22%	104/108 96,3%
s20,s22,s24	106/108 98,15%	106/108 98,15%	108/108 100%
s25,s27,s29	107/108 99,07%	108/108 100%	108/108 100%
s26,s28,s30	106/108 98,15%	106/108 98,15%	107/108 99,07%
s31,s33,s02	99/108 91,67%	101/108 93,52%	102/108 94,44%
s32,s01,s03	103/108 95,37%	101/108 93,52%	100/108 92,59%

In each test has 1080 reference models and 108 testings are done. Results of recognition accuracy in tables are given in two ways: as rational number in the form, number of correct recognitions / number of testings = number of correct recognition / 108, and in percentage value. By this way results of recognition accuracy are explicitly given in absolute and relative form. Absolute form give information about number of correct recognitions in observed testing and also more fully represents the accuacy of the speaker recognizer.

Three cases for square of amplitude characteristic of applied frequency selective ranges are observed in Tab. 1, “Exponential” as in Eq. (2) and “Sigmoidal” for parameter *a*=1.0 and *a*=0.5 as in Eq. (3). Recordings used for testing in appropriate test are listed in first column of tables. It is evident that in experiments for the same square of amplitude characteristic of applied frequency selective ranges, “Exponential” or “Sigmoidal”, recognition accuracy varies depending on which recordings are used for training and testing. Minimum of recognition accuracy is achieved when test set contains recordings {s02,s04,s06}. Sentences in this case are shorter compared with sentences in other test sets. Also recognition accuracy is smaller, below or around 90%, in experiments when test recordings are {s07,s09,s11} and {s01,s03,s05}. Other tests listed in Tab. 1, in which test sets do not contain recordings s02 or s03, show results of recognition accuracy higher than 95%. Test sets {s31,s33,s02} and {s32,s01,s03}, Tab. 1, contain shorter recordings s02 or s03 and achieved recognition accuracy is again smaller. Sentence s09 is also shorter sentence, it is of similar duration as sentence s04. The recognition accuracy for test set {s07,s09,s11} is below 90%.

In most tests application of sigmoidal square of amplitude characteristic of frequency selective ranges is increased recognition accuracy with respect to the case when this characteristic is exponential. Comparison of achieved mean recognition accuracies is given in Fig. 1. When square of amplitude characteristic of frequency selective ranges is exponential, Eq. (2), achieved mean recognition accuracy is approximately 94,14%. Application of sigmoidal square of amplitude characteristc, Eq. (3), increases mean recognition accuracy to approximately 94,75% for parameter *a*=1,0 and to approximately 94,98% for parameter *a*=0,5.

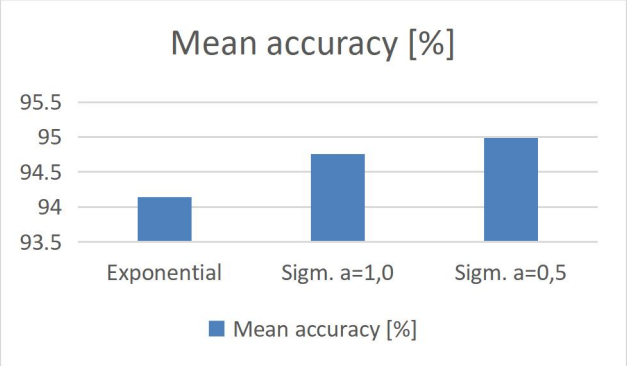

Figure 1 Mean recognition accuracies for columns of Tab. 1

Additional tests, Tab. 2, are done for previously mentioned test sets when recognition accuracy is smaller. When short test recording is replaced with longer test recording accuracy in that test is increased. The increase of accuracy is more pronounced if difference in duration of these recordings is larger.

Table 2 Results for feature vector 21 MFCCs, sigmoidal square of amplitude of frequency ranges, additionally applied sigmoid on elements of model, minimum and maximum accuracies in each of columns are shaded

Test set	Sigm. a=0.5	Sigmoid on model
s02,s04,s06	92/108 85,18%	99/108 91,67%
s02,s04,s20	96/108 88,89%	101/108 93,52%
s02,s06,s20	99/108 91,67%	103/108 95,37%
s07,s09,s11	97/108 89,81%	104/108 96,3%
s07,s09,s20	99/108 91,67%	104/108 96,3%
s07,s20,s24	105/108 97,22%	105/108 97,22%
s01,s03,s05	99/108 91,67%	105/108 97,22%
s01,s05,s20	105/108 97,22%	105/108 97,22%
s32,s01,s03	100/108 92,59%	106/108 98,15%
s32,s20,s03	101/108 93,52%	107/108 99,07%
s31,s33,s02	102/108 94,44%	105/108 97,22%
s31,s33,s20	106/108 98,15%	108/108 100%

For example, durations of test recordings of sentence s06 are larger than durations of test recordings s04, therefore the test set {s02,s06,s20} is recognized with higher recognition accuracy with respect to the test set {s02,s04,s20}. Duration of sentence s20 is larger than duration of sentence s06. Recognition accuracy for test set {s02,s04,s20} is higher than recognition accuracy for test set {s02,s04,s06}.

Recordings of sentences s20 and s24, as longer recordings, were used for replacement of shorter recordings. In some cases

replacement by the recording s20 is resulted in increasing of recognition accuracy higher than 5%. Sentence s20 is larger than sentence s11. Test set {s07,s09,s20} is recognized with higher accuracy with respect to test set {s07,s09,s11}. By replacing sentence s09 with larger sentence s24 it is achieved additional increase in recognition accuracy, from 91,67% for test set {s07,s09,s20} to 97,22% for test set {s07,s20,s24}, six recordings is additionally correctly recognized. Similar increasing of recognition accuracy is achieved when test set {s01,s03,s05} is replaced with test set {s01,s05,s20}, or when test set {s31,s33,s02} is replaced with test set {s31,s33,s20}.

Application of sigmoid function $sigm(x) = \frac{1}{1+e^{-x}}$ on elements of model i.e. on elements of appropriate covariance matrix, results by increasing of recognition accuracy (Tab. 2). These increasings of recognition accuracy are larger than increasings listed in Tab.1 achieved when square of amplitude characteristic of frequency selective ranges is changed from exponential to sigmoidal shape. In most tests, when initial recognition accuracy, accuracy without applied sigmoid function on elements of appropriate covariance matrices, is around 90% or less than 95%, increasing of recognition accuracy of around 5% is achieved.

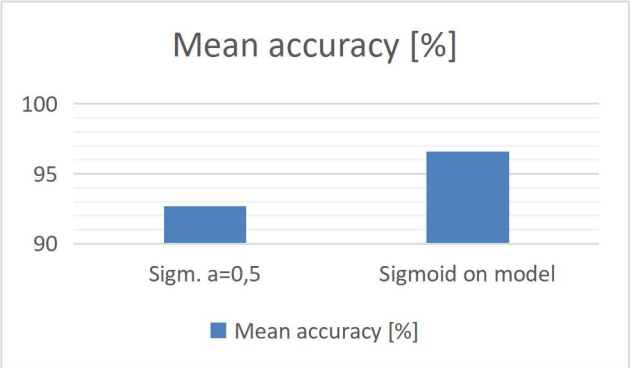

Figure 2 Mean recognition accuracies for columns of Tab. 2

By comparing achieved mean recognition accuracies for results presented in columns of Tab. 2 it is evident increasing of mean recognition accuracy from 92,67% to 96,6%, Fig.2. In this case, application of the sigmoid function to the elements of the used models results in an increase in the mean recognition accuracy of 3,93% approximately.

4 CONCLUSION

Recognition accuracy varies when training and test sets are varied. In future work will be interesting to analyze each test when speaker is not correctly recognized and try to find a way for correction of recognition, try to understand why speaker recognizer make wrong decision. Listed results point to three ways for increasing of recognition accuracy when MFCCs are used as short-term speaker features: careful selection of the training and test set, application of frequency selective ranges used for calculation of MFCCs whose shape of square of amplitude characteristic is sigmoidal, application of sigmoid function on elements of covariance matrices used for speaker modeling. Results in Tab.1 show that accuracy of automatic speaker recognition can have significant variations in dependence of the training and test set used in concrete experiment. By observing results from the first column of Tab. 1 when exponential square of amplitude characteristic of frequency selective ranges is used it is evident that minimum of recognition accuracy, of approximately 83,33%, is achieved for test set{s02,s04,s06} and that maximum of recognition accuracy, of approximately 99,07%, is achieved for test set {s25,s27,s29}. This is increasing of recognition accuracy of 15,74%, or it is increasing of 17/108. By analysing of the second and third column of Tab. 1 it is evident that difference between maximum and minimum achieved recognition accuracy in these columns is 100%-85,18%=14,82%. Also, by comparing maximum and minimum recognition accuracy in columns of Tab. 2 it is evident that difference between them in first column is 98,15%-85,18%=12,97% and in second column is 100%-91,67%=8,33%. These differencies between maximum and minimum achieved recognition accuracies for appropriate realizations of speaker recognizer have significant values and show that choice of the set for training and testing of speaker recognizer has a significant impact to recognition accuracy that can be achieved. Achieved maximums of recognition accuracies in columns of Tab. 1 and Tab. 2 show that for used set of sentences for training of the speaker recognizer there is a set of sentences for testing of speaker recognizer which maximizes recognition accuracy of speaker recognizer by achieving recognition accuracy between 98% and 100%. The opposite can also be argued, for choiced set of test sentences there is a set of training sentences which maximizes recognition accuracy between 98% and 100%.

The results demonstrate that integrating automatic speaker recognition using MFCCs within IoT frameworks can significantly contribute to security and optimization in smart cities, providing an effective tool for real-time monitoring, enhanced public safety, and efficient urban resource management.

Results in Tab. 2 and Fig. 2 show that if recognition accuracy initially not have sufficient value then accuracy can be increased by applying sigmoid function on elements of models. Sigmoid function has horizontal asymptotes, it is nonlinear function in limited domain. After applying of sigmoid function on numeric values, they are normalized and scaled in the limited range (0,1), recognition accuracy is significantly improved. In future work it can be experimented with ways of application of nonlinear functions, nonlinear functions that are based on exponential and sigmoid function, on numeric values used during automatic speaker recognition to achieve better recognition accuracy. Achieved accuracy for the test set {s02,s04,s06}, 91,67% in Tab. 2, indicates that here is necessary to apply additional transformation to achieve recognition accuracy higher than 95%, which is achieved in most of tests.

Results presented in this paper show that recognition accuracy of applied procedure for automatic speaker recognition depend of textual content of sets used for training and testing. This is consequence of the procedure which is used for automatic speaker recognition. To avoid textual dependency of accuracy of speaker recognizer it is necessary to much better extract information of color of voice from speech signal. MFCCs track the spectral envelope of speech signal which contains information about color of voice but spectral envelope is dependent of energy contained in speech signal. It means that spectral envelope depend of textual content of speech. Spectral envelope contains redundant informations with respect to automatic speaker recognition. Changes of this redundant informations in spectral envelope have impact to changes of MFCCs and cause wrong decision in procedure of automatic

speaker recognition. To extract information of color of voice from speech signal it is necessary to do adequate filtering of speech signal. Application of exponential or sigmoidal square of amplitude characteristic of frequency selective ranges which are used during calculation of MFCCs is one manner of filtering of speech signal to better extract information about color of voice.

5 REFERENCES

[1] Abdul, Z. Kh. and Al-Talabani A. K. (2022). Mel Frequency Cepstral Coefficient and Its Applications: A Review. *IEEE Access, 10*, 122136-122158. DOI: 10.1109/ACCESS.2022.3223444

[2] Ittichaichareon, C., Suksri, S., Yingthawornsuk, T. Speech Recognition Using MFCC. *In Proceedings of International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012),* Pattaya, Thailand; July 28-29, 2012, 135-138. DOI: 10.13140/RG.2.1.2598.3208

[3] Dhingra, S. D., Nijhawan, G., Pandit, P. (2013). Isolated Speech Recognition Using MFCC and DTW. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2*(8), 4085-4092.

[4] Elharati, H. A., Alshaari, M. and Këpuska, V. Z. (2020). Arabic Speech Recognition System Based on MFCC and HMMs. *Journal of Computer and Communications, 8*(3), 28-34. https://doi.org/10.4236/jcc.2020.83003

[5] Maurya, A., Kumar, D., Agarwal R.K. (2018). Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach. *6th International Conference on Smart Computing and Communications, ICSCC 2017,* 7-8 December 2017, Kurukshetra, India, *Procedia Computer Science, 125,* 880-887. https://doi.org/10.1016/j.procs.2017.12.112

[6] Devi, K. J., Devi, A. A., Thongam, K. (2019). Automatic Speaker Recognition using MFCC and Artificial Neural Network. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9*(1S), 39-42. DOI: 10.35940/ijitee.A1010.1191S19

[7] Wirdiani, A., Machetho, S. N., Putra, I. K. G. D., Sudarma, M., Hartati, R. S., Ferdian, H. A. (2024). Improvement Model for Speaker Recognition using MFCC-CNN and Online Triplet Mining. *International Journal on Advanced Science, Engineering and Information Technology, 14*(2), 420-427. https://doi.org/10.18517/ijaseit.14.2.19396

[8] Reggiswarashari, F., Sihwi, S. W. (2022). Speech emotion recognition using 2D-convolutional neural network. *International Journal of Electrical and Computer Engineering (IJECE), 12*(6), 6594-6601. http://doi.org/10.11591/ijece.v12i6.pp6594-6601

[9] Panda, B., Padhi, D., Dash, K., Prof. Mohanty, S. (2012). Use of SVM Classifier & MFCC in Speech Emotion Recognition System. *International Journal of Advanced Research in Computer Science and Software Engineering, 2*(3), 225-230.

[10] Attabi, Y., Alam, M. J., Dumouchel, P., Kenny, P., O'Shaughnessy, D. Multiple Windowed Spectral Features for Emotion Recognition. *Published in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing,* Vancouver, BC, Canada, May 26-31, 2013, 7527-7531. https://doi.org/10.1109/ICAS-SP.2013.6639126

[11] Bojanić, M., Delić, V., Sečujski, M. (2014). Relevance of the Types and the Statistical Properties of Features in the Recognition of Basic Emotions in Speech. *Facta Universitatis, Series: Electronics and Energetics, 27*(3), 425-433. https://doi.org/10.2298/FUEE1403425B

[12] Chou, C.-H., Ko, H.-Y. Automatic Birdsong Recognition with MFCC Based Syllable Feature Extraction. In: Hsu CH., Yang L.T., Ma J., Zhu C. (Eds.) *Ubiquitous Intelligence and Computing, UIC 2011, Lecture Notes in Computer Science,* Springer, Berlin, Heidelberg, 2011, 6905, 185-196. https://doi.org/10.1007/978-3-642-23641-9_17

[13] Cai, J., Ee, D., Pham, B., Roe, P., Zhang, J. Sensor Network for the Monitoring of Ecosystem: Bird Species Recognition. *Proc. 2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information,* Melbourne, VIC, Australia, 3-6 December, 2007, 293-298. https://doi.org/10.1109/ISSNIP.2007.4496859

[14] Zhang, S., Gao, Y., Cai, J., Yang, H., Zhao, Q., and Pan. F. (2023). A Novel Bird Sound Recognition Method Based on Multifeature Fusion and a Transformer Encoder. *Sensors* 2023*, 23*(19), 8099. https://doi.org/10.3390/s23198099

[15] Muhammad, G. and Alghathbar, K. (2013). Environment Recognition for Digital Audio Forensics Using MPEG-7 and Mel Cepstral Features. *The International Arab Journal of Information Technology, 10*(1), 43-50.

[16] Domazetovska, S., Gavriloski, V., Anachkova, M., Petreski, Z. (2021). Urban Sound Recognition Using Different Feature Extraction Techniques. *Facta Universitatis, Series: Automatic Control and Robotics, 20*(3), 155-165. https://doi.org/10.22190/FUACR211015012D

[17] Jokić, I., Delić, V., Jokić, S., Perić, Z. (2015). Automatic Speaker Recognition Dependency on Both the Shape of Auditory Critical Bands and Speaker Dicriminative MFCCs. *Advances in Electrical and Computer Engineering, 15*(4), 25-32. https://doi.org/10.4316/AECE.2015.04004

[18] Lyon, R. F., Katsiamis, A. G., Drakakis, E. M. (2010). History and Future of Auditory Filter Models. *Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS 2010),* May 30 – June 2 2010, Paris, France, 3809-3812. https://dx.doi.org/10.1109/ISCAS.2010.5537724

[19] Sigmund, M. (2019). Speaker Discrimination Using Long-Term Spectrum of Speech. *Journal of Information Technology and Control, 48*(3), 446-453. https://doi.org/10.5755/j01.itc.48.3.21248

[20] Cummins, F., Grimaldi, M., Leonard, T., Simko, J. The CHAINS Corpus: CHAracterizing INdividual Speakers. *In Proc. of the 11th International Conference "Speech and Computer" SPECOM'2006,* St. Petersburg, Russia, June 25-29, 2006, 431-435.

Contact information:

Ivan JOKIĆ, grades and ranks
(Corresponding author)
Year of birth: 1980
Institution: University Business Academy in Novi Sad, Faculty of Economics and Engineering Management in Novi Sad
Postal address: Cvećarska 2, 21107 Novi Sad, Srbija
Mail: ivan.jokic@fimek.edu.rs
https://orcid.org/0009-0008-0083-7675

Vlado DELIĆ, grades and ranks
Year of birth: 1964
Institution: University of Novi Sad, Faculty of Technical Sciences
Postal address: Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija
Mail:vlado.delic@uns.ac.rs
https://orcid.org/0000-0002-4558-9918

Zoran PERIĆ, grades and ranks
Year of birth: 1964
Institution: University of Niš, Faculty of Electronic Engineering
Postal address: Aleksandra Medvedeva 14, 18000 Niš, Srbija
Mail: zoran.peric@elfak.ni.ac.rs
https://orcid.org/0000-0002-8267-9541