# About Applicability of Automatic Speaker Recognition Algorithm for Automatic Recognition of Sounds in Nature

Ivan Jokić, Stevan Jokić, Vlado Delić, *Member, IEEE* and Zoran Perić, *Member, IEEE*

***Abstract*—This paper describe the Android application for recognition of arbitrary sounds. Application uses algorithm previously used for automatic speaker recognition. First 18 mel-frequency cepstral coefficients are used as sound features. Twenty exponential auditory critical bands are used during their determination. Covariance matrices are applied for modeling of sounds. User of this application creates own sound database. Application supports capture of sounds in mono wav format, 16 bit resolution, with frequency sampling of 22050 Hz. After sound capturing is finished user can choose between training and testing procedure. Training is devoted to creating of new model for observed sound and saving in sound database on memory card of mobile phone for example. Testing procedure run modeling of observed sound and compare that model with reference models in sound database. As result of recognition application writes the identity of most similar reference model and also the second and third most similar model and measures of difference with respect to them.**

***Index Terms*—Automatic speaker recognition, sound, automatic sound recognition, mel-frequency cepstral coefficients, covariance matrices, difference between models.**

## I. Introduction

SOUNDS are the constituent part of nature. Each source of sound makes his personal sounds. On that way each sound source can be differentiated with respect to others sources of sound. Therefore sound of each source can be used for recognizing of that source.

Development of technique enables that today human uses many technical devices in your everyday life, for example mobile phones, tablets, laptops or notebooks, desktop computers. All these technical devices are equipped with microphone and/or camera. Manufacturers of these devices tend to provide that users can communicate with their devices. For more naturalness interaction and communication between human and technical systems it is necessary that technical system can recognize the human, i.e. his identity. It means that technical system shall to have embedded algorithms for recognizing of individual person. On that way communication between the human and some technical system will be much more natural. So to have correct recognition from the side of technical system, technical system shall have complete audio visual information about human, but often systems for recognition are only based on audio or only on visual information. Often it is much faster to capture audio information about observed subject. Also, not infrequently only audio information about observed subject i.e. user of technical device are available. Implementation of algorithms for automatic speaker recognition enables that technical device recognize user by observing only his voice i.e. speech.

To achieve automatic speaker recognition it is necessary to define speaker features and modeling manner. Very popular are short time features, these features are based on analysis of short time intervals. For each short time interval appropriate feature vector is calculated. Mel-frequency cepstral coefficients (MFCCs) are one of often used speaker features [1], [2], [3], [4].

Today mobile phones, Android or iPhone, can process signals. This opens possibility for a lot of applications which can help the user to better understand nature around self and learn directly from nature. For example, it can be assumed that a lot of us are very amazed when hear song bird. But, not infrequently the problem is that we do not recognize which bird sings [5] [6], [7]. Also in nature human can meet many other sounds. Therefore it seems that developing of applications for recognizing of different sounds can be interesting for today human and improve his knowledge about nature. Since this problem is oriented towards recognizing of different timbres of sounds it is consequence that algorithm for automatic speaker recognition can be used.

The essence of algorithm for automatic speaker recognition i.e. used features and modeling manner, is applicable in many problems when the task is to recognize some information of interest from analyzed sound. In most problems when the task is some automatic recognition on sounds, in fact decision depends of spectral content of observed and analyzed sound. For example in speech recognition, different voices have different spectral content, in emotion recognition speech in different emotions have different spectral content, in speaker recognition speech of different speakers have different spectral content. MFCCs are consequence of spectral content

Ivan Jokić is with the University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 2100 Novi Sad, Serbia (e-mail: ivan.jokic@uns.ac.rs).

Stevan Jokić is with the ECG for Everybody (e-mail: stevan@ecg4everybody.com).

Vlado Delić is with the University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia, (e-mail: vdelic@uns.ac.rs).

Zoran Perić is with the University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia, (e-mail: zoran.peric@elfak.ni.ac.rs).

of observed speech segment and therefore it is not surprisingly that the same speech features, MFCCs, are used for automatic speech recognition [8], [9], [10], automatic emotion recognition from speech [11], [12], [13], and automatic speaker recognition. The main difference in these three areas of sound recognition is observed object of modeling, in speech recognition it is phoneme, in emotion recognition are emotions, and in speaker recognition are speakers.

## II. Features Used and Modeling Manner

Feature vectors are determined for sound frames of around 23 ms. Frames are shifted each other by around 8.33 ms. MFCCs are used as speaker features and they are calculated by equality [14]:

$$c_n = \sum_{k=1}^{20} \log(E_k)\cdot \cos\left[n\cdot\left(k-\frac{1}{2}\right)\right], \qquad n = 1,2,\ldots,18. \tag{1}$$

It is evident that determination of MFCCs is based on the estimated energies $E_k$ inside auditory critical bands. It was used 20 auditory critical bands, width of 300 mel which are mutually shifted by 150 mel. Based on the results in [15], where exponential shape of auditory critical bands is shown as more efficient with respect to rectangular and triangular shape, recognizer described in this paper also use exponential shape of auditory critical bands. Therefore energy inside appropriate auditory critical band, in (1) denoted by $E_k$ is determined by assumption that auditory critical bands are of exponential shape with square amplitude characteristic determined by:

$$A_{\exp}^2(k) = \begin{cases} e^{(k-k_{c,n})s}, & k_{1,n} \le k \le k_{c,n}, \\ e^{-(k-k_{c,n})s}, & k_{c,n} < k \le k_{2,n}, \end{cases} \tag{2}$$

where $k_{c,n} = \dfrac{k_{1,n}+k_{2,n}}{2}$ is a central discrete frequency of n$^{\text{th}}$ auditory critical band and $s$=2 is steepness factor.

Feature vectors of one source, one subject of interest or one speaker in fact, are mutually dependent. From the stochastic point of view in fact these feature vectors are distributed in accordance with some stochastic distribution. Since the sound production is natural process, it can be assumed, as is in many works, that MFCC feature vectors are distributed in accordance with appropriate Gaussian multidimensional distribution. Gaussian multidimensional distribution is described by vector of mean values and covariance matrix. Vector of mean values carries information about average values inside each dimension of observed feature vectors. Therefore vector of mean values depends of sample of feature vectors observed and as such can vary from sample to sample of speaker's speech. Opposite to that, covariance matrix in fact describes shape of feature vectors distribution and better describes timbre of voice or some sound in a general case. Therefore modeling was done by covariance matrices. During training and also during testing feature vectors are grouped into data matrix $X$. Appropriate model $\Sigma$ is calculated by:

$$\Sigma = \frac{1}{n-1}\cdot(X-\mu)\cdot(X-\mu)^T. \tag{3}$$

In (3) $n$ is number of feature vectors used for determination of model and $\mu$ is vector of mean values. Used algorithm makes decision of recognition based on comparison of test model and reference models in sound database. Measure of difference is defined by:

$$d(i,ref) = \frac{1}{18^2}\cdot\sum_{m=1}^{18}\sum_{n=1}^{18}\left|\Sigma_i(m,n)-\Sigma_{ref}(m,n)\right|. \tag{4}$$

During decision application determines difference between all reference models in sound database and model of observed sound. Identity of reference model for which difference is smallest is assigned to observed sound.

## III. Application Functionality

Application has three mode of working: recording, training and testing. It works on real recorded sounds and therefore in first step of application using it is necessary that user save target sound. User has at his disposal two buttons for this action: button "Start Recording" for start of record process and button "Stop Recording" for stopping of recording process. Sounds captured by this way are recorded into folder SoundRec in memory of mobile phone in file snimak.wav. Characteristics of the recorded sound are mono format in 16 bit resolution and with frequency sampling of 22050 Hz.

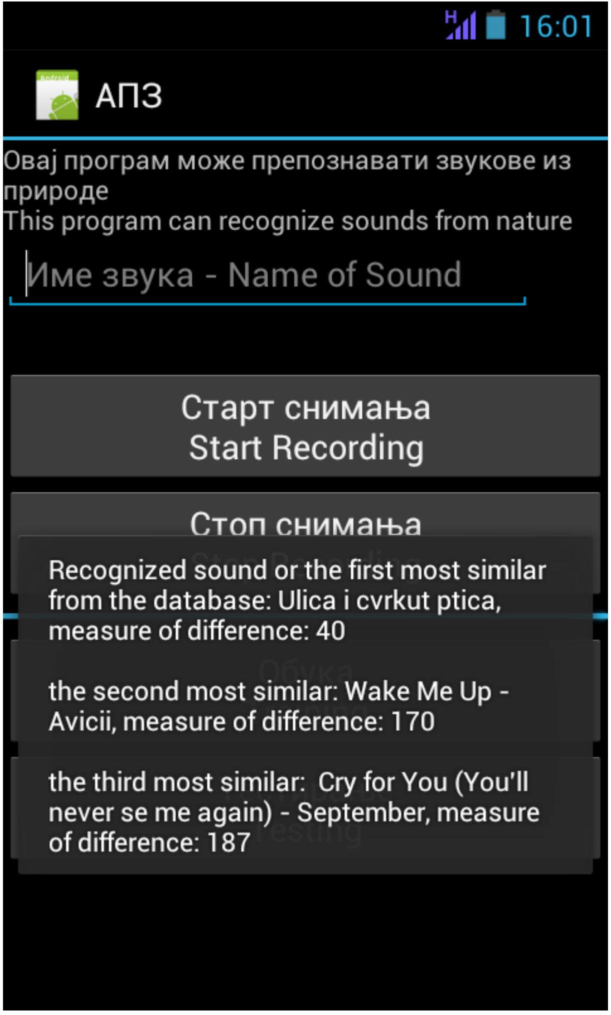
When sound is recorded, user can run training or testing procedure, buttons: Training and Testing. If user wants to add the model of sound from file snimak.wav in sound database then it is necessary to write into text box of application screen the name of sound and after that push the button Training. During training procedure application forms appropriate covariance matrix for sound from file snimak.wav and save this model into file modeli.txt, also the name of sound is saved into file names.txt. These two textual files are saved into same directory, SoundRec, of mobile phone, as the file snimak.wav.

User self creates your own sound database. When user is satisfied with the variety of models in sound database then he can start testing procedure for previously recorded sound in the file snimak.wav. In that case it is necessary that user push the button Testing, application starts forming of appropriate model for recording in file snimak.wav and compare this model with all models in sound database. As result of recognition application writes identity or name of sound in file snimak.wav. This identity or name is equivalent with the name of the most similar reference model in sound database.

Application accuracy depends of diversity of sound database. Since user self create sound database it can be case that user test on some sounds that seems as similar to some previously recorded sound which model is in sound database. In that case user can not be sure that most similar model represents identity of observed sound. In that case it is much more obvious that most similar model represent that the sound with that identity is the most similar sound to observed test sound, of course with respect to other models in sound database. So it seems that user can gain a better picture about unknown sound if he can have information about similarity of this sound to a few sounds in sound database which are most similar. Therefore in application was added possibility to show how much is total differentiation between observed test model and each of three most similar reference models. This differentiation between models is determined by:

$$d_T = \left\lfloor \left( \sum_{m=1}^{18} \sum_{n=1}^{18} \left| \Sigma_{test}(m,n) - \Sigma_{ref}(m,n) \right| \right) \right\rfloor . \qquad (5)$$

This equation is very similar to (4), with respect to (4) in (5) is not done normalization and as result the positive integer value of differentiation is observed. This is done for better understanding of results on the side of the user of application.



Овај програм може препознавати звукове из природе
This program can recognize sounds from nature

Recognized sound or the first most similar from the database: Ulica i cvrkut ptica, measure of difference: 40

the second most similar: Wake Me Up - Avicii, measure of difference: 170

the third most similar: Cry for You (You'll never se me again) - September, measure of difference: 187

Fig. 1. Screenshot of mobile phone when application writes results of recognition.

Application can be used in many different cases. In sound database user can save different sounds, sounds of different nature. These sounds can be, for example, some bird's song or sound, or sounds of some animals, but for example, these sounds can be also parts of songs, parts of played songs. This case, when sound database consists of sounds of different nature is presented on Fig. 1. This is only one example of using of this application when formation of sound database is being started. In moment of testing shown on Fig. 1 sound database was consisted only of few sound models and some of models correspond to most known parts of some popular songs. Test was done on sound recorded in the university campus to capture song of some bird. Therefore recognized model i.e. the most similar model is in accordance with test sound, it is model of song bird captured from the little street with very little traffic, this reference model is named in Serbian as "Ulica i cvrkut ptica", translated in English it can be named as "Street and song bird". As results show the measure of difference is the smallest, it is 40 and it is much less with respect to second and third most similar models. Also it is evident that measures of difference with respect to second and third most similar model in fact are similar i.e. of the same order of magnitude in this case of testing, it is 170 and 187 respectively. This can be explained as a consequence of the fact that the second and third most similar model are models of played songs which are very different with respect to sound captured in recording with bird song.

The user can choose to record just sounds of one nature. These can be for example only sounds of birds or only parts of songs or some sounds of some different nature. On that way user in fact uses this application as automatic recognizer of for example sounds of birds or automatic recognizer of songs or as automatic recognizer of sounds of some other nature.

## IV. Conclusion

It is evident that this application is applicable in many different cases. Also it can be expected that this application give to the end user useful and interesting services. User can create his own sound database. On that way user can create a lot of reference models. Displayed measure of difference during testing can be used as help to the user to conclude whether the model for recorded sound adds or not to sound database. Application gives the user information about difference with respect to three most similar models, and this measure is as something absolute measure. Therefore based on the value of differentiation, if differentiation is large, user can conclude to add model of recording in sound database since that model is something new for sound database in that moment. On that way user refresh his sound database.

Sometimes it can be expected that user can be interested to have in sound database some sounds which he can not to capture. In this case user is interested to have measures of difference of test model with respect to reference models. This can be solved if for example some most popular sound databases will be hosted on some place on Internet. In that case user could download these sound databases from appropriate Internet site. User can use this sound database as finished sound database. Also user could refresh this sound database with some new models and on that way expand the sound database.

## ACKNOWLEDGMENT

This work is result of our work on the following projects: TR32035, TR32032, DANSPLAT (E!9944), CABUNS.

## REFERENCES

[1] T. Kinnunen, H. Lee, “An overview of text-independent speaker recognition: From fetures to supervectors,” *Speech Communication*, vol. 52, no. 1, pp. 12-40, 2010.

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Margin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz, and D. A. Reynolds, “A Tutorial on Text-Independent Speaker Verification,” *EURASIP Journal on Applied Signal Processing* 2004:4, pp. 430-451, 2004.

[3] M. M. Dobrović, V. D. Delić, N. M. Jakovljević, I. D. Jokić, “Comparison of the Automatic Speaker Recognition Performance over Standard Features,” in Proc. of the 2012 IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics (SISY 2012), Subotica, Serbia, pp. 341-344, September 20-22, 2012, [Online]. Available: http://dx.doi.org/10.1109/SISY.2012.6339541

[4] V. Tiwari, “MFCC and its applications in speaker recognition,” *International Journal on Emerging Technologies*, vol. 1(1), pp. 19-22, 2010.

[5] CH. Chou, HY. Ko, “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, vol. 6905, pp. 185-196. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-23641-9_17

[6] M. Deepika, A. Nagalinga Rajan, “Automatic Identification of Bird Species from the Recorded Bird Song Using ART Approach,” *IJIRSET*, vol. 3, Special Issue 3, pp. 668-675, Mar., 2014.

[7] J. Cai, D. Ee, B. Pham, P. Roe, J. Zhang, “Sensor Network for the Monitoring of Ecosystem: Bird Species Recognition,” Proc. *2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information*, Melbourne, Qld., pp. 293-298, 3-6 Dec., 2007. DOI: 10.1109/ISSNIP.2007.446859, URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4496859&isnumber=4496790

[8] C. Ittichaichareon, S. Suksri, and T. Yingthawornsuk, “Speech Recognition using MFCC,” in Proc. International Conference on Computer Graphics, Simulation and Modeling (ICGSM’2012), Pattaya (Thailand), pp. 135-138, July 28-29, 2012.

[9] S. D. Dhingra, G. Nijhawan, P. Pandit, “Isolated speech recognition using MFCC and DTW,” *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 2, issue 8, pp. 4085-4092, August 2013.

[10] S. V. Arora, “Effect of Time Derivatives of MFCC Features on HMM Based Speech Recognition System,” *International Journal on Signal and Image Processing*, vol. 4, no. 3, pp. 50-55, September 2013. DOI: 01.IJSIP.4.3.1257

[11] D. Neiberg, K. Elenius and K. Laskowski, “Emotion Recognition in Spontaneous Speech Using GMMs,” in *INTERSPEECH 2006 – ICSLP*, Pittsburg, Pennsylvania, pp. 809-812, September 17-21, 2006.

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

[13] Y. Attabi, M. J. Alam, P. Dumouchel, P. Kenny, D. O’Shaughnessy, “Multiple windowed spectral features for emotion recognition,” Published in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, pp. 7527-7531, May 26-31, 2013.

[14] B. R. Wildermoth, “Text-Independent Speaker Recognition Using Source Based Features,” pp. 19-20, M. Phil. Thesis, Griffith University, Brisbane, Australia, January 2001.

[15] I. Jokić, V. Delić, S. Jokić, Z. Perić, “Automatic Speaker Recognition Dependency on Both the Shape of Auditory Critical Bands and Speaker Discriminative MFCCs,” *Advances in Electrical and Computer Engineering*, vol. 15, no. 4, pp. 25-32, November 2015, doi:10.4316/AECE.2015.04004.