



Use of Covariance Matrix in Automatic Speaker Recognition

Ivan JOKIĆ¹, Stevan JOKIĆ²

Abstract: One procedure for automatic speaker recognition based on use of 21 mel-frequency cepstral coefficients as speaker features and covariance matrix as speaker model is tested in this paper. Tests are conducted on the Solo part of the CHAINS speech database which contains 37 recordings for each of 36 speakers. Each speech recording is represented by appropriate matrix of feature vectors. Modeling of recording of speaker is done by covariance matrix of matrix of feature vectors. Results of recognition accuracy are compared for two cases, when on elements of speaker model is applied sigmoid function and when it is not. Tests are done in five stages. Application of sigmoid function on elements of covariance matrices results in most of tests in significantly increasing of recognition accuracy. Achieved mean recognition accuracy for all done tests when sigmoid function is not applied is 87,84% and when sigmoid function is applied is 94,64%.

Keywords: Automatic speaker recognition; Mel-Frequency Cepstral Coefficients; Covariance matrix.

1 INTRODUCTION

In connections of different devices over the Internet and enabling voice commands it is necessary to achieve recognition of speaker's voice. Voice has three basic characteristics: intensity, basic frequency and color. Color of voice is consequence of harmonic structure of spectrum of voice. Information about intensity and basic frequency are also contained in spectrum of voice signal. Envelope of spectrum of voice signal contains information about harmonic structure of voice.

Short-term analysis of speech signal is based on analysis of speech frames of around 25 ms mutually shifted by around 10 ms. Mel-frequency cepstral coefficients (MFCCs) are usually used as short-term features of speech signal. They are used as short-term features for automatic speaker recognition [1]. To implement speaker recognition system in [2] MFCCs are combined with Gaussian mixture models (GMMs), in [3, 4] MFCCs are combined with artificial neural network. MFCCs track spectral envelope of speech signal and therefore contain information implied by spectral envelope which is observed. Therefore MFCCs contain information about color of voice i.e. about identity of speaker, but also they contain information about textual content of speech and also about emotional state of speaker. In [5] MFCCs are used for speech recognition, recognition of textual content of speech, and in [6, 7] MFCCs are used for recognition of emotion in speech.

Considering sound in general, MFCCs can be used as features of sounds from the environment [8] or from nature, for bird sound recognition [9] and for classification of honey bees [10] for example. As is evident from literature, MFCCs are used in many applications oriented to recognitions different information from sound [11].

Modeling is the step in automatic speaker recognition after short-term features calculation. Model make compact picture of previously calculated short-term features. It is calculated for considered set of short-term features. Model describes longer intervals in sound, in fact, model is long-term feature [12]. Proces of automatic speaker recognition begin with short-term features and also combine long-term features [13], to recognize identity of speaker i.e. information of interest from speech.

Models of the same speaker should be as similar as possible. Similarity can be increased by applying adequate transformation on elements of model. In this paper will be presented that application of sigmoid transformation on model is additionally increased recognition accuracy. In next part of paper is described speaker recognizer used and speech database which is used. After that results of recognition accuracy are presented.

2 DESCRIPTION OF EXPERIMENTAL SETUP

MFCCs are used as short-term speaker features. They are calculated by:

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

where $n=\{1, 2, \dots, 21\}$ is ordinal number of calculated MFCC. Feature vector contains 21 MFCC. E_m is energy inside of m -th frequency selective range:

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

Frequency range of interest is divided into $M=22$ frequency selective ranges, width of 300 mels and mutually shifted by 150 mels. $X(k)$ is discrete Fourier transform (DFT) of observed speech signal frame $x(n)$, $k_{1,m}$ is lower discrete frequency and $k_{2,m}$ is upper discrete frequency of observed m -th frequency selective range. Each frame of speech signal is windowed by Hann window function:

$$w(n) = \frac{1}{2} \cdot \left(1 - \cos \frac{2 \cdot \pi \cdot n}{N-1} \right), 0 \leq n \leq N-1, N=1024. \quad (3)$$

The frequency sampling of observed speech signals is 44100 Hz. Usually used duration of speech frame is around 25

ms. To apply the fast Fourier transform algorithm to calculate the discrete Fourier transform of the frame of observed speech signal it is necessary that duration of frame is number which can be represented in the form of 2^s , where s is a natural number. In accordance with this, the frame duration was chosen to be 1024 points, i.e. approximately 23 ms. Neighboring frames are shifted by 368 samples, approximately by 8ms. Square of amplitude characteristic of applied frequency selective ranges is of sigmoid shape [14], given by equation:

$$A_m(k)^2 = \begin{cases} \text{sigm}_1(k - k_{c,m}), & k_{1,m} \leq k < k_{c,m}, \\ 1, & k = k_{c,m} \\ \text{sigm}_1(k_{c,m} - k), & k_{c,m} < k \leq k_{2,m}. \end{cases} \quad (4)$$

In equation $k_{1,m}, k_{2,m}$ and $k_{c,m} = (k_{1,m} + k_{2,m})/2$ are the lower, upper and central discrete frequency of m -th frequency selective range. Shape of sigmoid function used is determined by equality

$\text{sigm}_1(x) = \frac{1}{1 + e^{-0.5 \cdot x}}$. For each speech frame is calculated feature vector of 21 MFCCs. Feature vectors of the recording of one speaker are written in appropriate columns of matrix of feature vectors.

Covariance matrix of matrix of feature vectors is used for modeling of speaker:

$$\Sigma = \frac{1}{n-1} \cdot (X - \mu) \cdot (X - \mu)^T, \quad (5)$$

$[X]_{d \times n}$ -matrix of feature vectors, d -dimensionality of feature vector used i.e. number of MFCCs used, $d=21$, n -number of feature vectors, $[\mu]_{d \times 1}$ -vector of mean values of matrix X . Each element in covariance matrix observe appropriate energy representation of MFCCs. Diagonal elements represents energy in appropriate dimensions of feature vectors while elements outside of the main diagonal of the covariance matrix describe the measure of correlation between appropriate dimensions of feature vectors. Measure of difference between two models, for example of model of test speech and reference model, is calculated by equality:

$$r(\Sigma_{test}, \Sigma_{ref}) = \frac{1}{d^2} \cdot \sum_{i=1}^d \sum_{j=1}^d |\Sigma_{test}(i, j) - \Sigma_{ref}(i, j)|. \quad (6)$$

The identity of most similar reference model with respect to Eq. (6) is recognized speaker. If in speech database have R reference models then the test speech has identity of the i -th reference model if $r(\Sigma_{test}, \Sigma_i) < r(\Sigma_{test}, \Sigma_j), j \in \{1, 2, \dots, R\} \setminus \{i\}$.

Experiments are conducted on the part of the speech database CHAINS (CHAracterizing INdividual Speakers) [15] named Solo. This part of the used speech database is characterized by property – subjects simply read a prepared text at a comfortable rate. CHAINS speech database contains recordings of speech of 36 speakers. In the Solo part, for each of the 36 speakers, there are 37 recordings: four short fables {f01, f02, f03, f04} and sentences {s01, s02, ..., s33}. Recordings are in WAV format, frequency sampling is 44100 Hz and quantisation resolution is 16 bit/sample.

3 RESULTS OF RECOGNITION

Testings of automatic speaker recognition are done in five steps. In first four experiments for training are used recordings from the set of short fables, f01, f02, f03 and f04. Testings are conducted in 12 tests, in each of tests appropriate set of pronounces of three sentences is used as test set, Tab. 1 and Tab. 2. In the fifth experiment, Tab. 3, the 33 recordings from set of sentences {s01, s02, ..., s33} are used for training and testing of recognizer. Tests are organized in 12 tests as in Tab. 1 and Tab. 2. In each of 12 tests, training is done on the rest of 30 recordings which are not used in observed test set.

Two cases are compared, when as model of speaker is used covariance matrix calculated by Eq. (5) and when sigmoid function $\text{sigm}(x) = \frac{1}{1 + e^{-x}}$ is applied on elements of previously calculated model i.e. covariance matrix. Elements of model vary in time. Sigmoid function is a nonlinear, companding, function. Values of sigmoid function are limited in the range [0, 1]. Therefore it can be expected that application of sigmoid function on elements of model will decrease time variation of elements of models of the same speaker. This function is usually used as activation function in artificial neural networks.

Table 1 Training: f01, f02

Test set	Cov. matrix		Sigmoid applied	
	f01	f02	f01	f02
s01,s02,s03	75%	82,41%	88,89%	88,89%
s04,s05,s06	70,37%	74,07%	80,56%	82,41%
s07,s08,s09	72,22%	70,37%	93,52%	89,81%
s10,s11,s12	86,11%	86,11%	97,22%	94,44%
s13,s14,s15	83,33%	86,11%	91,67%	94,44%
s16,s17,s18	87,96%	86,11%	92,59%	93,52%
s19,s20,s21	88,89%	91,67%	91,67%	94,44%
s22,s23,s24	80,56%	82,41%	92,59%	87,96%
s25,s26,s27	91,67%	90,74%	95,37%	96,30%
s28,s29,s30	84,26%	83,33%	95,37%	95,37%
s31,s32,s33	84,26%	90,74%	95,37%	99,07%
s20,s31,s33	87,04%	93,52%	95,37%	99,07%

Recognition accuracy varies depending of test set. Applying of sigmoid transformation on model in Tab. 1 is resulted in increasing of recognition accuracy in all experiments. Recognition accuracy when training is done with recording f01 is smaller than 90% for all test sets except for test set {s25, s26, s27}. By applying sigmoid transformation on elements of model recognition accuracy in certain tests is higher than 95%. When f02 is used for training recognition accuracy in some tests is higher than recognition accuracy achieved when f01 is used for training. Application of sigmoid transformation in experiments when test sets are {s31, s32, s33} and {s20, s31, s33} increases recognition accuracy to 99%. Significant increasing of recognition accuracy is evident in Tab. 1 when test set is {s07, s08, s09}. In this case when f01 is used for training increasing is higher than 20% and when f02 is used for training increasing is higher than 19%. Results when f03 and f04 are used for training, Tab. 2, are similar to results in Tab. 1. For training done by f03 here are two cases, for test set {s16, s17, s18} and {s19, s20, s21}, when applying of sigmoid function on model is not increased recognition accuracy. In some cases when recognition accuracy is below or around 80%, application of

sigmoid transformation increases recognition accuracy around or higher than 10%. The smallest value of recognition accuracy for training done by f03 or f04 is for test set {s07, s08, s09}. Increasing of recognition accuracy after applying sigmoid function on elements of model is around 15%.

Table 2 Training: f03, f04

Test set	Cov. matrix		Sigmoid applied	
	f03	f04	f03	f04
s01,s02,s03	85,18%	86,11%	88,89%	94,44%
s04,s05,s06	78,7%	78,7%	83,33%	88,89%
s07,s08,s09	76,85%	74,07%	91,67%	91,67%
s10,s11,s12	92,59%	89,81%	93,52%	98,15%
s13,s14,s15	87,04%	90,74%	92,59%	95,37%
s16,s17,s18	94,44%	89,81%	93,52%	97,22%
s19,s20,s21	93,52%	92,59%	93,52%	96,3%
s22,s23,s24	89,81%	85,18%	92,59%	93,52%
s25,s26,s27	93,52%	91,67%	98,15%	97,22%
s28,s29,s30	87,04%	87,96%	97,22%	98,15%
s31,s32,s33	90,74%	91,67%	100%	100%
s20,s31,s33	93,52%	92,59%	100%	100%

Recognition accuracy when training and test set are from set of sentences {s01, s02, ..., s33}, Tab. 3, and when test sets are from sets {s04, s05, s06} or {s07, s08, s09} is significantly higher than in cases when training recordings are f01, f02, f03 or f04. This indicate that recognition accuracy depend of textual content of training and test speech. After application of sigmoid function on elements of model when test sets are {s01, s02, s03} and {s07, s08, s09} accuracy of recognition is increased to the value higher than 95%.

Table 3 Training and test on the set {s01, s02, ..., s33}

Test set	Cov. matrix	Sigmoid applied
s01,s02,s03	88,89%	95,37%
s04,s05,s06	87,96%	91,67%
s07,s08,s09	89,81%	95,37%
s10,s11,s12	97,22%	99,07%
s13,s14,s15	94,44%	97,22%
s16,s17,s18	97,22%	100%
s19,s20,s21	99,07%	100%
s22,s23,s24	98,15%	99,07%
s25,s26,s27	99,07%	100%
s28,s29,s30	100%	99,07%
s31,s32,s33	97,22%	100%
s20,s31,s33	98,15%	100%

Table 4 Mean recognition accuracy

Training	Cov. matrix	Sigmoid applied
f01	82,64%	92,51%
f02	84,8%	92,98%
f03	88,58%	93,75%
f04	87,58%	95,91%
s01,...,s33	95,6%	98,07%
mean accuracy	87,84%	94,64%

Mean recognition accuracies for different training recordings used, Tab. 4, show that application of sigmoid transformation on elements of covariance matrices is significantly increased accuracies. In the case when training is done by recording f01, the increase of mean recognition accuracy is the largest, approximately 10%. The largest mean recognition accuracy is achieved when training and test recordings are from the set of sentences {s01, ..., s33}, in this case mean recognition accuracy is approximately 98%. Mean

accuracy for all tests, when used model is covariance matrix, is 87,84%. When sigmoid transformation is applied on elements of models this mean accuracy is increased to 94,64%. Application of sigmoid function on elements of models is decreased mean error of recognition from around 12% to around 5%.

4 CONCLUSION

Sigmoid transformation applied on elements of covariance matrix of matrix of MFCC feature vectors increases recognition accuracy. The recognition results confirm that application of sigmoid function on elements of model decreases time variation of the elements of models of the same speaker. From results in Tab. 1 and Tab. 2 for test sets {s01, s02, s03}, {s04, s05, s06}, {s07, s08, s09}, it is evident that when the recognition accuracy is less than 80%, application of sigmoid transformation on elements of model can increase recognition accuracy above 80%. In some cases increased accuracy is close and around 90% or higher than 90%, but there is also cases when increased accuracy is below 85%. In Tab. 3 recognition accuracy for these three test sets is higher than 87% and accuracy after applying of sigmoid transformation is higher than 91%. It is evident that recognition accuracy depends of training and test set.

Result of recognition of automatic speaker recognizer should just depend of the characteristics of the voice. Decision of recognizer which is described in this paper depend of textual content of training and test speech. In future work it is necessary to find a way that decision do not depend of textual content of training and test speech. One of solutions can be by using appropriate filters during calculation of MFCCs and also by using appropriate transformations during calculation of models i.e. long-term features of speaker which will decrease time variation of elements of models of the same speaker.

Acknowledgements

Thanks to Professor Vlado Delić from the Faculty of Technical Sciences, University of Novi Sad, and Professor Zoran Perić from the Faculty of Electronic Engineering, University of Niš.

5 REFERENCES

- [1] Kinnunen, T., Li, H. (2010). An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. *Speech Communication*, 52(1), 12-40. <https://doi.org/10.1016/j.specom.2009.08.009>
- [2] 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 (2018), 880-887. <https://doi.org/10.1016/j.procs.2017.12.112>
- [3] 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. <https://doi.org/10.35940/ijitee.A1010.1191S19>
- [4] 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/ijaset.14.2.19396>

- [5] Elharati, H. A., Alshaari, M. and Kępuska, 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>
- [6] 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>
- [7] 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>
- [8] 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>
- [9] 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>
- [10] Libal, U., Biernacki, P. (2024). MFCC-Based Sound Classification of Honey Bees. *International Journal of Electronics and Telecommunications*, 70(4), 849-853. <https://doi.org/10.24425/ijet.2024.152069>
- [11] Abdul, Z. Kh. and Al-Talabani A. K. (2022). Mel Frequency Cepstral Coefficient and Its Applications: A Review. *IEEE Access*, 10, 122136-122158. <https://doi.org/10.1109/ACCESS.2022.3223444>
- [12] 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>
- [13] Büyük, O., Arslan, M. L. (2018). Combination of Long-Term and Short-Term Features for Age Identification from Voice. *Advances in Electrical and Computer Engineering*, 18(2), 101-108. <https://doi.org/10.4316/AECE.2018.02013>
- [14] Jokić, I., Delić, V., Perić, Z. Application of Mel-Frequency Cepstral Coefficients in Automatic Speaker Recognition as Part of IoT Solutions for Security and Optimization in Smart Cities. *ALFATECH Journal*, 1(1), in press.
- [15] 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: Cvečarska 2, 21207 Novi Sad, Srbija

K-Mail: ivan.jokic@fimek.edu.rs

<https://orcid.org/0009-0008-0083-7675>

Stevan JOKIĆ, grades and ranks

Year of birth: 1983

Institution: Alfa BK University, Faculty of Information Technologies

Postal address: Bulevar maršala Tolbuhina 8, Novi Beograd, Srbija

L-Mail: stevan.jokic@alfa.edu.rs

<https://orcid.org/0000-0003-4432-0172>