Baseline system for speaker recognition - parameter analysis

Željko Nedeljković, Milana Milošević, Željko Đurović

Abstract— **Interest for speaker recognition increased again with appearance of new opportunities for applications. The goal of this paper is to examine reliable speaker recognition system and to examine influence of different parameter values. The system consists of Gaussian Mixture Model (GMM) based classifier which uses Mel Frequency Cepstral Coefficients (MFCC) as features. Tests were conducted on Temporal Voice Idiosyncrasy (TEVOID) database. First we performed test using parameters recommended in literature and established the baseline system with excellent 99.5% of recognition rate. We examined the varieties in following steps of process: in speech preprocessing, MFCC calculation, GMM training and in training/testing setup. All the tests were performed using only 10 sentences in total for training and testing. The results of these tests are close to saturation, which makes them not robust enough for general conclusions, but useful for further development.**

***Index Terms*—MFCC; GMM; Speaker recognition**

I. Introduction

The topic of speaker recognition was extensively investigated during past few decades. The main goal of speaker recognition is to identify who is speaking, as reliable as it is possible, using a sample of speech. In last few years, interest in this topic has raised again as new applications have appeared. Modern speaker recognition applications demand deeper analysis of speech signal. Addressing the issue of emotionally colored speech is one which is interesting for the authors. Here are some examples of such tasks. Companies which have call centers for serving the large number of customers are interested in fast and reliable recognition of callers in order to cut call times and costs. This becomes a challenging task when caller is not in neutral emotional state. Next, there is analysis of turn speaking or “who spoke when” in multiple party conversations [1]. In this case, vivid discussion among speakers will be a real challenge. Also, forensic researches are always a hot topic.

As a first step to build a system for speaker recognition with taking into consideration emotionally colored speech, we implement standard speaker recognition system as a part of SEBAS [4] framework on the neutral speech database. The goal of this work is to verify reliability of the system on simple task and to analyze influence of different parameter values on total recognition rate when only 10 sentences are used both for training and testing. System consists of the Gaussian Mixture Model (GMM) [2] classifier using Mel Frequency Cepstral Coefficients (MFCC) [3] as features. All tests were performed on Temporal Voice Idiosyncrasy (TEVOID) database [5, 6]. The parameters we analyze are configuration parameters for MFCC and GMM, different proportion of test/training data and different lexical content of test/training data on total recognition rate.

The reminder of this paper is following: In Section II we briefly describe the TEVOID. In Section III we describe testing speaker recognition system – GMM and MFCC. Next, in section IV results are presented and commented. Final conclusion and plans for feature investigation are presented in section V. In the end, we give the review of literature

II. The TEVOID Database

The TEVOID [5, 6] as the database is particularly designed to study speech temporal variability across a highly homogeneous group of speakers. 50 speakers - 24 male and 26 female are fluent native speakers of the same language variety and the same age group (between 20 and 30). All speakers are fluent native speakers of Zurich German. For each speaker 256 read sentences of neutral speech were recorded. The sentences were long and short, phonetically representative. All the data were recorded at 44.1 kHz sampling rate [5, 6].

III. Speaker recognition system

Speaker recognition system which we analyze in this paper is GMM [2] classifier using MFCC [3] as features. The basic information on MFCC [3] extraction from speech signal, and GMM training is given in following.

A. MFCC

MFCC [3] is the most widely used speech feature. These coefficients represent an audio signal based on human perception. MFCC [3] is calculated according to the following procedure: the speech signal is divided into frames, then Hamming window is applied and after that each frame is transformed into the frequency domain using DFT. Next, a bank of normalized triangle filters, equally spaced on the mel-scale, is employed. Discrete cosine transformation and logarithm are then applied to the filter output to obtain the mel-frequency cepstrum (MFC). The procedure is illustrated

Željko Nedeljković is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: nz135003p@student.etf.bg.ac.rs).

Milana Milošević is with the School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra, 11020 Belgrade, Serbia (e-mail: mm125026p@student.etf.bg.ac.rs).

Željko Đurović – Signals & Systems Division, School of Electrical Engeenering, University of Belgrade, Kralj Aleksandar’s Bulevard 73, 11020 Belgrade, Serbia (e-mail: zdjurovic@etf.bg.ac.rs).

on Figure 1.

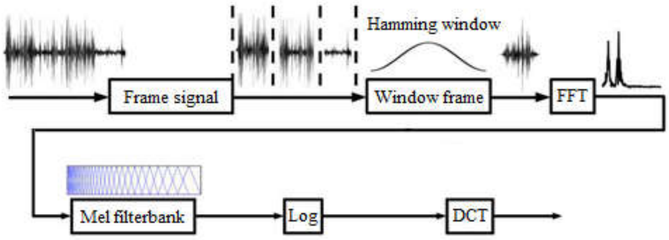

Fig. 1. MFCC extraction from speech signal block diagram

B. *GMM*

A GMM [2] speaker model consists of finite number of mixtures of multidimensional Gaussian probability density function. Expectation maximization algorithm [7] is used for maximizing the likelihood with respect to given data. For algorithm initialization one iteration k-means clusterization of training data was used [2]. Minimal value in diagonal covariance matrix was limited to 0.01.

IV. Results and Discussion

We examined influence of parameterization in different stages of speaker recognition task: MFCC [3] feature extraction, Gaussian mixture model [2] training and finally in training/test setup. The results are presented in following sections.

The baseline system was tuned using parameter values recommended by literature [2, 3, 8] as guidelines. The baseline setup is given in Table 1. In addition to that, in the baseline setup we didn't use signal normalization, nor delta and double delta coefficients. The baseline system performed as high as 99.5%. This is at some point expected since the database is of high quality regarding speaker differentiation as shown in pervious experiments on this database [9]. Since this result is close to saturation limit, it doesn't leave much space for in depth parameter analysis.

TABLE I
Baseline Parameters

	Parameter	Value
MFCC	frequency range	300-3700Hz
	num. channels	20
	coefficients	0-12
GMM	num mixtures	30
setup	num sentences train	6
	num sentences test	4

From the baseline setup, parameters were systematically examined one by one, in reasonable range of values [1]. To examine influence of particular parameter, we were changing the value of only that parameter, while other parameters of the system had baseline values. The results of the experiments are given in following subsections.

A. *Signal preprocessing*

In this step, we examined influence of speech-only extraction from signal and signal mean value bias. Using speech extraction from utterances and by eliminating pauses recognition rate went up to 100% . Since the baseline result was also very high, we can consider that there is no much difference for this database with or without extraction of speech-only part of utterance. In the case of mean value removal, unexpectedly, the recognition rate lowered to 92.5%.

B. *Training/testing setup*

We limited the number of sentences used for model training and testing to 10 all together. This decision is made due to limits in real-world applications and data available for feature research.

a) Number of sentences for test and training

Experiments started using only one sentence per speaker for training and the rest for testing, and then increased the number of sentences for training and decreased the number of sentences for testing.

The results are shown on the Figure 2.a. It was interesting to check, how the system will behave if the number of test sentences is always the same – 10. The results of this experiment are lower than the previous one as expected (Figure 2.b).

b) Different sets of sentences

To check the influence of lexical content of the speech, we examined 10 by 10 sets of sentences from database. Configuration 6 sentences for training and 4 for test were used. The results are shown on the Figure 3.

This test demonstrated that, in this setup, lexical content of the speech almost doesn't have impact on recognition rate. For that reason, we decided to continue all the experiments on the first 10 sentences.

C. *MFCC feature extraction*

As far as MFCC parameter analysis is concerned, we examined different range of MFCC and different number of coefficients, different number of channels in MFCC calculation, and frequency range.

a) Different range of MFCC

In this test we examined the influence of different coefficients of the MFCC used for feature vector. The length of feature vector was 13. Sets of coefficients used were 0th-12th, next 1st-13th,…, 7th-19th. The results are shown on Figure 4. The results confirm that the first coefficients are the most appropriate for speaker recognition task.

a) Number of MFCCs

In this experiment, we examined influence of different number of MFCC included in feature vector, starting from 1 (only 0th) to 20 (0th-19th). The results are show on Figure 5. It can be noted that there is only slight variation in recognition rate for using more than first 9 coefficients (including 0th).

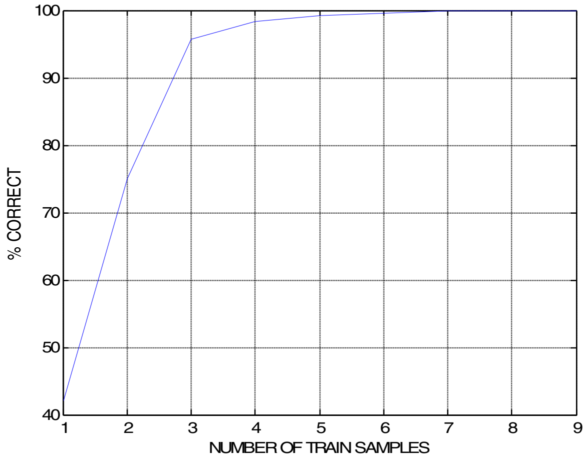

Fig. 2.a. Recognition rate depending on the number of training sentences when number of training and test sentences used in total is 10.

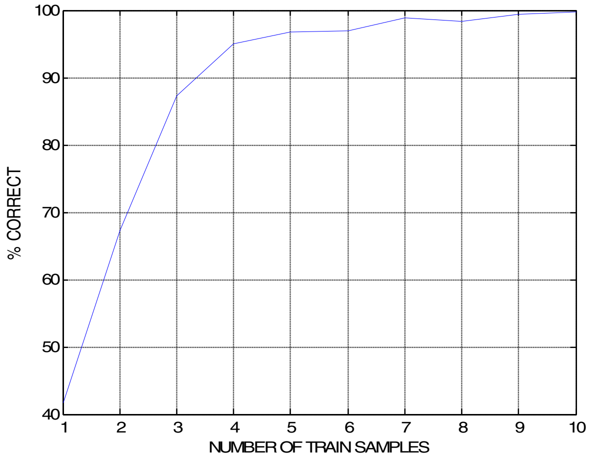

Fig. 2.b. Recognition rate depending on the number of training sentences when number of test sentences is fixed to 10.

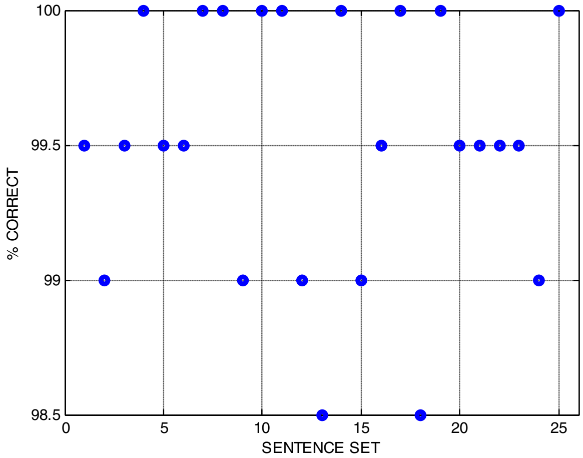

Fig. 3. Recognition rate depending on set of sentences used

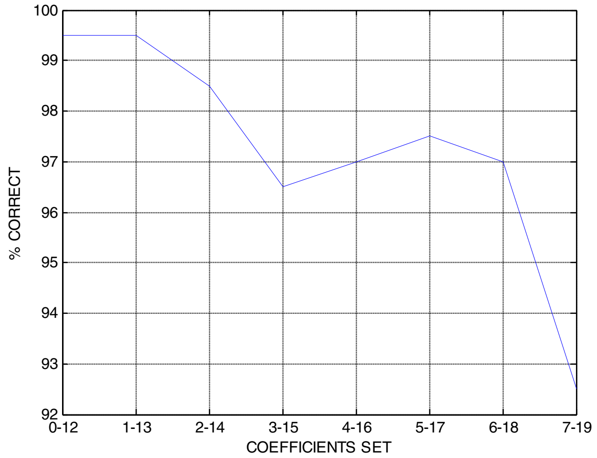

Fig. 4. Recognition rate depending on coefficient set used.

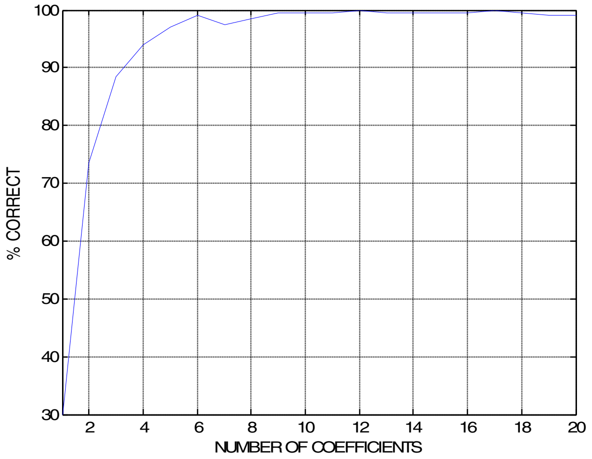

Fig. 5. Recognition rate depending on the number of MFCC coefficients used.

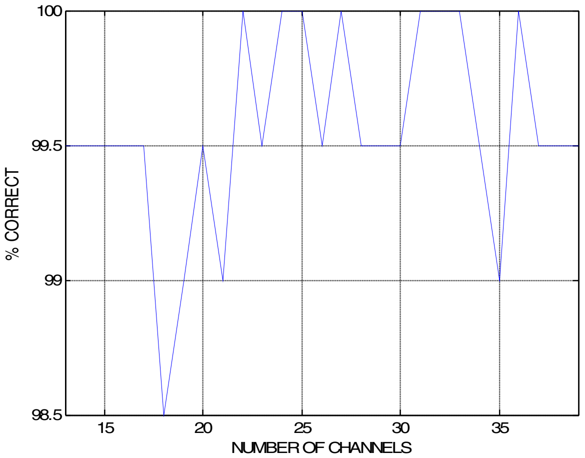

Fig. 6. Recognition rate depending on number of channels for MFCC calculation

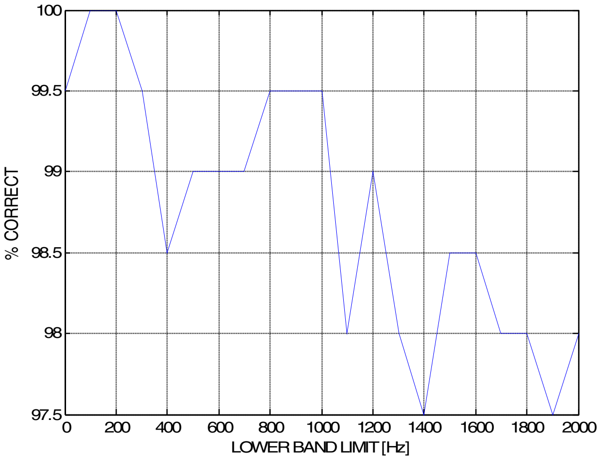

Fig.7. Recognition rate depending on lower band limit for MFCC calculation.

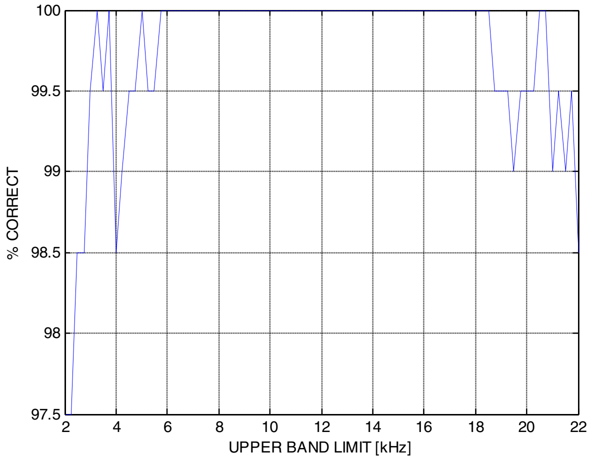

Fig. 8. Recognition rate depending on upper band limit for MFCC calculation

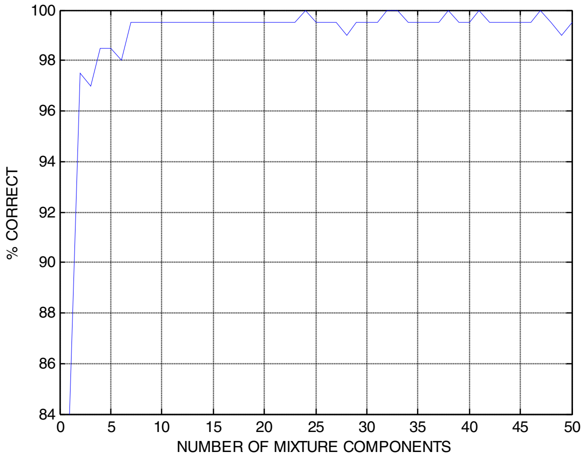

Fig. 9. Recognition rate depending on the number of Gaussian components in model.

b) Number of channels

Tests started with triangular filter bank which contains only 13 training filters, which is minimal number considering that we have 13 coefficients, and then increased the number of channels up to 40. The results are shown on Figure 6. According to these results, the desirable number of channels is more than 21 and less than 35.

c) Frequency band

The lower frequency limit didn't have any impact until 1000Hz when the performance dropped significantly. As far as upper limit is concerned, the results obtained for upper limit over 3500Hz are on saturation limit and more or less the same. The results indicate that optimal choice of upper limit would be from 6000Hz to 18000Hz. The results are shown on Figures 7 and 8.

d) Deltas

By adding deltas in feature vector of baseline system, recognition rate rised up to 100%.

D. Gaussian mixture model training

We were changing the number of Gaussian Mixture Model components used for a single speaker model – from 1 to 50 and the resulting recognition rate is plotted on Figure 9. We concluded that 8 mixtures is enough in this case.

V. Conclusion

The tests performed in this paper showed how speaker recognition rate changes with change of system parameters when database of neutral speech – TEVOID is in use. The system using this setup and data is robust to single parameter change as far as values of parameters are sensible.

The lexical content of TEVOID database doesn't have much influence on speaker recognition rate. As far as number of sentences for training/testing is considered, using more than 5 sentences for training gives acceptable recognition rates.

Next, the first set of MFCC coefficients, the best result are obtained when coefficients from 0^{th} to 12^{th} are used. Also, the test for number of coefficients showed that increasing the number of coefficients doesn't have an impact after first 9 coefficients. One more parameter we examined in MFCC calculation was the number of channels. Although value of this parameter doesn't have significant influence on recognition rate, it is desirable to be between 21 and 35. Finally, recognition rate decreases with the increase of lower frequency limit above 1000Hz, while upper frequency limit didn't have a noticeable impact above 3500Hz.

In the end, number of mixtures in GMM gave stable and high results after 8 mixtures used in model.

In the future work, we aim to investigate influence of emotional speech on speaker recognition task and to explore possibilities for using this system and results for more robust speaker recognition.

References

[1] Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. Speech communication, 52(1), 12-40.

[2] Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE transactions on speech and audio processing, 3(1), 72-83.

[3] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing, 28(4), 357-366.

[4] Milošević, M., Nedeljković, Ž., & Đurović, Ž. (2016) SVM Classifier for Emotional Speech Recognition in Software Environment SEBA. 3rd International Conference on Electrical, Electronic and Computing Engineering IcETRAN 2016, Zlatibor, Serbia, June 13 – 16, 2016

[5] Dellwo, V., Leemann, A., and Kolly, M.-J. (2012). "Speaker idiosyncratic rhythmic features in the speech signal," in Proceedings of INTERSPEECH 2012, Portland, pp. 1582-1585.

[6] Leemann, A., Kolly, M. J., & Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. Forensic science international, 238, 59-67.

[7] Bishop, C. M. (2006). Pattern recognition. Machine Learning, 128, 1-58.

[8] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., ... & Valtchev, V. (2002). The HTK book. Cambridge university engineering department, 3, 175.

[9] Milošević, M., Glavitsch, U., Dellwo V.& He L. (2016). Segmental features for automatic speaker recognition in a flexible software framework. The 25th annual conference of the International Association for Forensic Phonetics and Acoustics, at York, United Kingdom, July 24 – 27, 2016