Introduction of Improved UWB Speaker Verification System*

Aleš Padrta and Jan Vaněk

University of West Bohemia, Department of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
{apadrta, vanekyj}@kky.zcu.cz

Abstract. In this paper, the improvements of the speaker verification system, which is used at Department of Cybernetics at University of West Bohemia, are introduced. The paper summarizes our actual pieces of knowledge in the acoustic modeling domain, in the domain of the model creation and in the domain of score normalization based on the universal background models. The constituent components of the state-of-art verification system were modified or replaced by virtue of the actual pieces of knowledge. A set of experiments was performed to evaluate and compare the performance of the improved verification system and the baseline verification system based on HTK-toolkit. The results prove that the improved verification system outperforms the baseline system in both of the reviewed criterions – the equal error rate and the time consumption.

1 Introduction

The speaker recognition task is being solved for several years at the Department of Cybernetics. Many experiments with the constituent components of the verification system were performed since that time. At first, the module designed for acoustic modeling [5] was tested. Then, the module, which ensures the training of the model [6] was investigated. Finally, the main verification module, which evaluates the correspondence of the test data and the target speaker model [7], was tested. The constituent components were based on techniques, which were originally designed for the speech recognition [4][8] and they were modified for the speaker verification purposes.

The components, which are borrowed from the speech recognition systems, cannot be further modified, because all meaningful possible modification has been already done. Thus, we have to propose the alternative modules, which are designed primary for speaker recognition. This allows a better adaptation of the modules to the speaker recognition task and further improve the performance of the speaker recognition system.

* The work was supported by the Ministry of Education of the Czech Republic, project no. MSM 235200004, and by the Grant Agency of the Czech Republic, project no. 102/05/0278.

This paper introduces the components, which are primary designed for the speaker recognition task. The baseline verification system based on HTK and the improved verification system based on new components are introduced in Section 2. The experiments are described in Section 3 and their results are discussed in Section 4. Finally, a conclusion is given in Section 5.

2 Description of Verification Systems

2.1 Feature Vectors, Acoustic Modelling

The acoustic modeling module has to extract suitable speaker characteristics, which allow us to distinguish the individual speakers. These characteristics are further used for the model training and the following verification. Both tested speaker verification systems are based on the short-time spectral characteristics. A non-speech events detector is used in both systems, but they differ in the detector working domain.

Baseline System The baseline features are computed by the HTK-Toolbox. The features are standard Mel-Frequency Cepstral Coeficients (MFCC) augmented by delta and acceleration coefficients. A preemphasis coefficient is 0.97. The length (resp. an overlap) of a Hamming window is 25 (resp. 15) millisecond. A Mel-frequency filter bank contains 25 triangular filters. Then, 13 cepstral coefficients, including zero-th (log-energy) coefficient, are computed. 13 delta and 13 acceleration coefficients are added. The final dimension of the feature vector is 39. A voice activity detector [4] removes non-speech segments from input wave files, i.e. it works in the time domain.

Improved System The schema of the signal processing module is showed in Figure 1. It is a modified version of MFCC extended by a voice activity detector. An input speech signal is preemphasised with the coefficient 0.97. The Hamming window has 25 millisecond length and 15 millisecond overlap. A power spectrum is computed by FFT. 25 triangular band filters are set up linearly in the mel-scale between 200 and 4000 Hz. The logarithms of band-filters outputs are decorrelated by the discrete cosinus transformation (DCT). The computed cepstrum has 20 coefficients without the zero-th (log-energy) coefficient which is discarded. A time sequence of the each coefficient is smoothed by a 11 frames long Blackman window. Then the delta coefficients are added. The final dimension of the feature vector is 40. A downsampling with factor 3 is applied to the final features for the reduction amount of data. At the end, frames, which were marked as non-speech event, are removed. The non-speech event detector estimates the noise level and the speech level independently in each band. If the estimated speech level is lower than the estimated noise level then the actual feature vector is marked as non-speech event and is discarded.

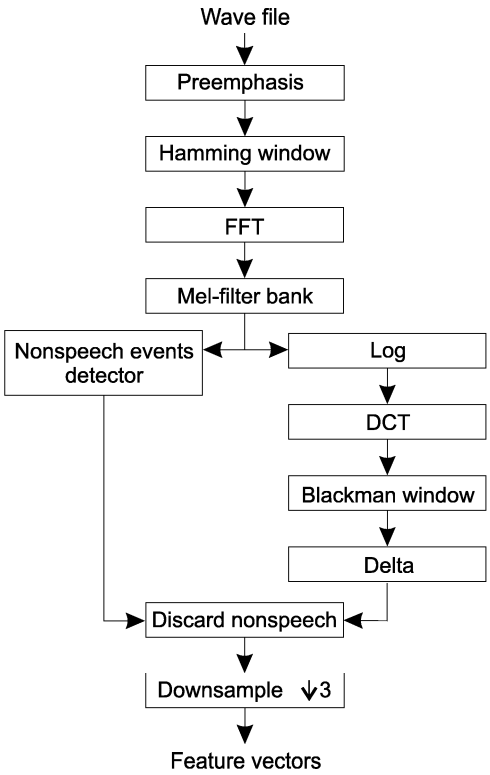

Fig. 1. Schema of acoustic modeling for improved system.

2.2 Speaker Models

The speaker model, created by the model-training module, has to represent the training data exactly. Next, a good ability of data generalization is desired, because the training set is limited. The short-time spectral characteristics are usually modeled by Gaussian mixture model (GMM) [1]. The baseline system and the improved system are based on GMM, but they differ in the model training techniques.

Gaussian Mixture Model A Gaussian mixture density of a feature vector $\boldsymbol{o}$ given the parameters λ is a weighted sum of M component densities, and is given by the equation

$$p(\boldsymbol{o}|\lambda) = \sum_{i=1}^{M} c_i p_i(\boldsymbol{o}), \tag{1}$$

where $\boldsymbol{o}$ is an N-dimensional random vector, $p_i(\boldsymbol{o})$, $i = 1, \ldots, M$, are the component densities, and c_i, $i = 1, \ldots, M$, are the mixture weights. Each component

density is an N-variate Gaussian function of the form

$$p_i(\boldsymbol{o}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{(\boldsymbol{o}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{o}-\boldsymbol{\mu}_i)\right\} \tag{2}$$

with the mean vector $\boldsymbol{\mu}_i$ and the covariance matrix $\boldsymbol{\Sigma}_i$. The mixture weights satisfy the constraint

$$\sum_{i=1}^{M} c_i = 1. \tag{3}$$

The complete Gaussian mixture density model is parameterized by the mean vectors, the covariance matrices, and the mixture weights from all component densities. These parameters are collectively represented by the notation

$$\lambda = \{c_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}, \quad i = 1, \ldots, M. \tag{4}$$

Baseline System The baseline system used the HTK-toolbox for GMM training. This HTK-toolbox was originally designed for the training of Hidden Markov models (HMM). It can be used for GMM training, because an one-state self-loop continuous density HMM (Figure 2) is equivalent to a GMM. As mentioned above, the HTK-toolbox was designed for the speech recognition and its training procedure (Expectation-Maximization algorithm) is optimized for this purpose. The model created by this optimized EM algorithm differs from the model created by the standard EM algorithm.

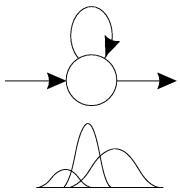

Fig. 2. One-state self-loop CDHMM: Schema of HMM and appropriate output probability density modeled by GMM.

Improved System We desire a robust model training procedure, which is able to create the model from a relative small amount of the training data, and a high precision of the model as well. First condition is fulfilled by the Distance-based (DB) algorithm [2], but its precision is not quite accurate due to the interpretation of the clusters as the mixtures of GMM. Thus, the model created by the DB algorithm was used as the initial condition for the standard EM algorithm. The combination of the DB algorithm and the EM algorithm results in a stable, fast, and precious training procedure even for a small amount of the train data [6].

2.3 Verification Algorithm

The verification algorithm is the same for the baseline verification system and the improved verification system. The goal of the speaker verification systems is to determine whether a given utterance is produced by the claimed speaker or not. This is performed by comparing a score, which reflects the agreement between the given utterance and the model of the claimed speaker, with an a priori given threshold. In verification systems based on GMM the simplest score is the likelihood of the utterance given the model of the claimed speaker.

Assume that there is a group of J reference speakers and that each speaker is represented by a Gaussian mixture model. We denote the model of the j-th speaker as ρ_j, $j = 1, \ldots, J$. Further suppose that a test utterance O consists of I feature vectors $\boldsymbol{o}_i$, $i = 1, \ldots, I$. The score reflecting the agreement between the feature vector $\boldsymbol{o}_i$ and the speaker j is then represented by the likelihood $p(\boldsymbol{o}_i|\rho_j)$ and is computed according to the formula (1).

However, such a score is very sensitive to variations in text, speaking behavior, and recording conditions, especially in the utterances of impostors. The sensitivity causes wide variations in scores, and makes the task of the threshold determination a very difficult one. In order to overcome this sensitivity, the use of the normalized score based on a background model has been proposed [1]. The score is then determined as the normalized log likelihood $\overline{p}(\boldsymbol{o}_i|\rho_j, \Lambda)$,

$$\overline{p}(\boldsymbol{o}_i|\rho_j, \Lambda) = \log(p(\boldsymbol{o}_i|\rho_j)) - \log(p(\boldsymbol{o}_i|\Lambda)), \tag{5}$$

where $p(\boldsymbol{o}_i|\Lambda)$ is the likelihood of the background model computed again using the formula (1).

3 Experimental Setup

3.1 Speech Data

Both verification system were tested on three different databases. An overview of the databases and their properties is given in Table 1. A detail description follows.

Table 1. Overview of the databases used for tests.

	UWB_S01	YOHO	KING
Number of speakers	100	138	51
Number of models	100	552	51
Number of test data	100	5,520	51
Number of trials	10,000	16,560	2,601

A part of the UWB_S01 corpus [3] was used in our experiments. It consists of speech of 100 speakers (64 male and 36 female). In our experiments, 40 sentences

per speaker were used from close-talking microphone. Further, the utterances were downsampled to 8 kHz and divided into two parts: 20 utterances of each speaker were used for the training of the GMMs of the reference speakers, and one other utterance of each speaker was used for the tests. Further, 2 other utterances of each speaker were reserved for the training of the background model for the first half of speakers (the speakers with the indexes 1-50). Each model was verified with each test utterance, i.e. the total number of the performed trials was $100 \cdot 100 = 10,000$.

The recordings from the corpus YOHO [9] were used for the next test of the verification systems. It consists of a speech of 138 speakers. The training data were recorded in 4 sessions and each was modeled independently, i.e. $138 \cdot 4 = 552$ models were created. The test data were recorded in 10 sessions, each of them was tested independently. 5 other sentences from each train session for the speakers 1-69 were reserved for the background model training. Each model was verified with all target speaker sessions and 20 impostor sessions, which were randomly chosen. It means, $552 \cdot 10 + 552 \cdot 20 = 16,560$ trials were performed.

Third testing database was the KING[10] database. Only a part of the recordings, acquired via the narrow-band microphone, was used in our experiments. The utterances were divided into two parts: 5 utterances of each speaker were used for the training of the GMMs of the reference speakers and one other utterance of each speaker was used for the tests. Further, one other utterance of each speaker was reserved for the training of the background model for the first half of speakers (the speakers with the indexes 1-25). Each model was verified with each test utterance, i.e. the total number of the performed trials was $51 \cdot 51 = 2,601$.

3.2 Description of Experiments

The tests of the performance and the time consumption were measured for the baseline verification system and for the improved verification system. Both criteria were tested on the three above specified databases. The performance of the systems was measured by the equal error rate (EER). The measured time represents the duration of all steps, which are necessary to evaluate all of the specified trials: the acoustic modeling of the train data, the test data, and the background model data; the training of the speaker models and the background model; the verification procedure.

The number of the mixtures was set to 32 for all speaker models and to 256 for the background models. These settings were common for all performed experiments. The training procedures are described in Section 2.2. The number of re-estimations was set to 9 in case of the baseline verification system and to 16 in case of the improved verification system [6].

4 Experimental Results

In Table 2, the performance of both systems are presented for the above mentioned databases. The first column identifies the name of the database, second

column contains EER values of the baseline verification system and the last column contains the results of the improved verification system. It can be seen from Table 2 that the improved verification system outperforms the baseline system in all tested databases.

Table 2. Overview of the EER for various databases for both systems

Corpus	HTK-based baseline System	Improved System
UWB_S01	1.00%	0.97%
YOHO	1.83%	1.72%
KING	16.20%	13.46%

The time consumed in the experiments is presented in Table 3. The structure of the table is the same as in case of Table 2, but the duration of the tests is showed instead of the EER values. The time data are in hh:mm format. We can say after the inspection of the results in Table 3 that the improved verification system needs significantly less time to perform the speaker verification than the baseline verification system based on the HTK-toolbox. The time savings are mainly in the model training module [6].

Table 3. Overview of the consumed time for various databases for both systems

Corpus	HTK-based baseline System	Improved System
UWB_S01	9:19	0:17
YOHO	23:17	1:08
KING	6:17	0:16

5 Conclusion

In this paper, an improved verification system was introduced. This system consists of the modules primary designed for the speaker recognition, i.e. it does not use modified components, which were originally designed for speech recognition. At first, the acoustic modeling module, which incorporates the non-speech events detector, was presented. Then, a model training module, which is capable of a fast model creation even from a relative small amount of the training data, was described. The proposed improved verification system was compared with our baseline verification system based on the HTK-toolbox. Three databases were used to evaluate the performance of both systems. The results show that the improved system need less time to perform the verification than the baseline system. Furthermore, the performance of the improved system is better than the performance of the baseline system. It can be said, that the proposed system based on the new modules outperforms of the baseline system.

References

1. Reynolds, D. A.: Speaker identification and verification using Gaussian mixture speaker models. Speech Communication **17** (1995) 91–108
2. Zilca, R. D., Bistricz, Y., "Distance-Based Gaussian Mixture Model for Speaker Recognition over the Telephone", ISCLP 2000, pp. 1001-1003, 2000
3. Radová, V., Psutka, J.: UWB_S01 Corpus – A Czech Read-Speech Corpus. Proc. ICSLP 2000 Beijing China (2000) 732–735
4. Prcín, M., Müller, L., Šmídl, L.: Statistical Based Speech/Non-speech Detector with Heuristic Feature Set. SCI 2002 – World Multiconference on Systemics, Cybernetics and Informatics Orlando FL-USA (2002), pp. 264–269
5. Vaněk, J., Padrta, A.: Optimization of features for robust speaker recognition, In Speech processing. Prague : Academy of Sciences of the Czech Republic, 2004. pp. 140-147. ISBN 80-86269-11-6.
6. Vaněk, J., Padrta, A., Radová, V.: An algorithm for speaker model training, Eurospeech 2005, Lisabon, 2005 (submitted)
7. Padrta, A., Radová, V.: On the background model construction for speaker verification using GMM, TSD 2004 - Text, speech and dialogue, Berlin: Springer, 2004. pp. 425-432. ISBN 3-540-23049-1. ISSN 0302-9743.
8. S. Young, et al.: The HTK book , Cambridge, 1995.
9. Campbell, J. P.: Testing with the YOHO CD-ROM voice verification corpus, ICASSP 1995, 1995
10. Campbell Jr., J. P., Reynolds, D. A.: Corpora for the Evaluation of Speaker Recognition Systems, ICASSP 1999, 1999