Feature Vector Classification by Threshold for Speaker identification

Sang-min Yoon, Kyung-mi Park, Jae-Hyun Bae, Yung-hwan Oh

Division of Computer Science Department of EECS
Korea Advanced Institutes of Science and Technology
{sangmin, kmpark, jhbae, yhoh}@speech.kaist.ac.kr

Abstract

This paper describes a new feature vector classification method for speaker identification. Purpose of this paper is constructing robust speaker models which only use meaningful feature vectors and discard confusing feature vectors. To construct robust speaker model, proposed method classifies feature vectors using log-likelihood estimation. Experimental results, with various segments ranging from 0.5 to 5 s, showed that our method outperforms previous method.

1. Introduction

Speaker identification is one of the major research area which uses speaker recognition technology. Speaker identification system selects the speaker who has the highest likelihood when the test utterance is given [1]. In text-independent speaker identification system, Gaussian Mixture Model (GMM) is the most frequently used likelihood function for speaker modeling. Conventional GMM based system is composed of two phases, the training phase and the test phase. The purpose of the training phase is constructing one speaker model for every speaker with feature vectors from each speaker's training data. In the test phase, maximum likelihood estimation is performed with constructed speaker models [2]. In general, feature vectors from different speakers sometimes contain similar characteristics because of acoustic similarity between speakers, background silence, and environment noises [4]. Because of the similar feature vectors, overlap of speaker models which lowers the system accuracy is created.

Recently, a feature vector classification method for robust speaker identification was proposed [4]. The proposed method classifies feature vectors from each speaker into two categories, overlapped and non-overlapped. Then, separated feature vectors are used to reconstruct two speaker models, an overlapped model and a non-overlapped model, for each speaker. In the test phase, only the feature vectors which have the maximum probability with non-overlapped speaker model are used to identify a speaker and others are discarded. As a result, the influence of overlapped feature vectors is decreased and fairly high accuracy is guaranteed when speaker models are heavily overlapped.

However, this method has some drawbacks. In the training phase, they didn't consider the reason why feature vectors are overlapped when they classifying them into two categories. If there are more overlapped feature vectors than non-overlapped ones and most of them are caused by the acoustic similarity between speakers, the accuracy will be lowered than conventional method due to the lack of feature vectors. In this paper, a new feature vector classification method is proposed to classify useful overlapped feature vectors from overlapped feature vector pool using log-likelihood estimation.

* This work was partially supported by Defense Acquisition Program Administration and Agency for Defense Development under the contract.

This paper is organized as follows : Section 2 explains the conventional GMM based speaker identification. Section 3 describes the speaker identification system based on selective use of feature vectors and proposes the new feature vector classification method. Experiments and results are showed in Section 4. Section 5 concludes this work and gives some future research works.

2. Conventional speaker identification

Speaker identification is a part of pattern recognition area. Like other pattern recognition works, speaker identification is composed of two phase, the training phase and the test phase.

The purpose of the training phase is constructing one speaker model for every speaker. First, feature vectors are extracted from training data. Mel-Frequency Cepstral Coefficient (MFCC) is one of the widely used speech spectral features. Second, maximum likelihood model parameters are estimated using the iterative expectation-maximization (EM) algorithm. GMM is known to be the most successful likely-hood function for text-independent speaker identification. In general, GMM for each speaker is constructed in training phase [2,3].

In the test phase, constructed models are used to identify test utterances. Under the assumption of independent feature vectors, the log likelihood of a model λ_i, (i =1,… ,S where there are S speakers) for a sequence of feature vectors X = $\{\vec{\mathbf{x}}_1,\ldots,\vec{\mathbf{x}}_T\}$, which are extracted from test data, is computed as follows:

$$\log p(X|\lambda_i) = \frac{1}{T}\sum_1^t \log p(\vec{x}_t|\lambda_i) \qquad (1)$$

Then, model $\hat{\mathbf{i}}$ which has the maximum probability with the given test utterance is chosen.

In this approach, the decision is critically depends on the interrelation between speaker models. Specially, overlapped regions of probability density functions (or speaker models) contribute to decision errors. Usually, the more speakers are in speaker identification system, the bigger overlapped regions get. Hence it is important to mitigate the overlap effects.

3. Feature vector classification by threshold for speaker identification

Background silence, environment noise, and acoustically similar features among speakers are known to the causes of overlapped regions between speaker models. Background silence and environmental noises are common features which are generally contained in a data stream of every speaker. Because of the influence of such common features, each feature vector set from different speakers may contain similar feature vectors. As a result, the discrimination power is lowered. Hence it is important to reduce the effect of the overlapped regions caused by common features.

However, overlapped features caused by acoustically similar feature between speakers should not be dumped. As mentioned above, common features are presented in every speaker's

feature space. But acoustically similar features would be presented in the feature space of some speakers who have the similar voice characteristics. Because those features can contribute to make right decision, they should be classified from other overlapped features. In this paper, we focus on classifying overlapped features into two categories, which are common features and acoustically similar features, not to drop useful features.

3.1 Robust speaker identification based on selective use of feature vectors

Recently, a new method which classifies feature vectors into two categories was proposed [4]. Feature vectors are classified into overlapped and non-overlapped categories in that system. First, they construct a speaker model for every speaker just like conventional speaker identification. With the constructed speaker models, maximum likelihood calculation is performed for every feature vector from training utterances. There could be some feature vectors falsely recognized if competing speaker models are overlapped. These misrecognized feature vectors are classified into overlapped category and the others are classified into non-overlapped category. After feature vector classification, the system constructs two models for each speaker. Non-overlapped speaker models are constructed with non-overlapped feature vectors and overlapped speaker models are constructed with overlapped feature vectors.

In the test phase, a sequence of feature vectors is extracted from test utterance. Then maximum likelihood calculation is performed for every feature vectors with reconstructed models. There will be some feature vectors which have the maximum probability with non-overlapped speaker model and the others will not. The system dumps the latter feature vectors and performs maximum likelihood calculation with feature vectors which were identified non-overlapped feature vectors.

This method can be useful for sequentially identifying speakers when speaker models are heavily overlapped. The more overlap gets bigger, the more feature vectors may contribute to decision error. By using this method, the speaker identification system can use only robust feature vectors to make right decision.

However, the accuracy of this system can be lowered than conventional method in some cases. It is because that this method dumps every overlapped feature vectors without regarding to the cause of overlap. In case of short utterance identification, we can get only limited number of feature vectors. When there are more overlap feature vectors than non-overlapped ones and most of them are caused by acoustic similarity, we can use only a few number of feature vectors to select a speaker. When more speakers are enrolled, this problem can significantly lower the system performance. In this paper, we try to overcome this problem to use the meaningful overlapped feature vectors.

3.2 Feature vector classification by threshold for speaker identification

In this chapter, we propose our feature vector classification method to use the overlapped feature vectors which can contribute to discrimination power. Fig. 1 shows the diagram of the feature space. Each circle represents one speaker's feature space. Gray colored area represents non-overlapped region, dark-gray colored area represents overlapped region caused by acoustic similarity and black colored are represents overlap region caused by common features.

Assume that a feature vector x_j is laid on the overlapped region caused by acoustically similar feature between two conventional speaker models. We made hypotheses as follows.

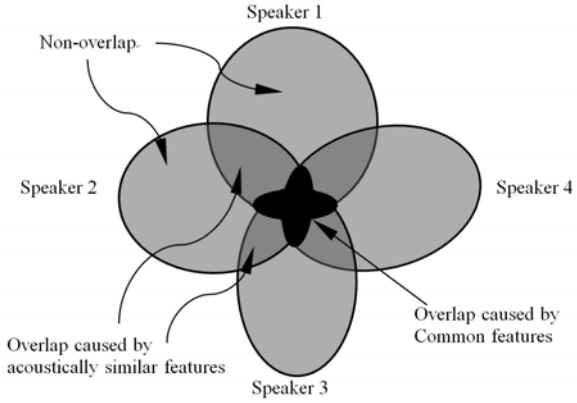


Figure 1: Diagram of feature space

First, the difference between two likelihood values with each competing model will be small. Second, the probability with the correct speaker model will be quite higher than the average probability of every speaker model. If an overlapped feature vector doesn't satisfy these two conditions, it may be laid on the overlapped region caused by common features.

Assume there are S conventional speaker models M_i (where i = 1,…,S). A feature vector x_j was used to construct speaker model M_s, and has the maximum probability with speaker model $M_{\hat{s}}$, where $\hat{s} = \text{argmax}\ log\ p(x_j|M_i)$.

If x_j is on the overlapped region caused by acoustically similar features between speakers, then following expression should be true when T_1 is a threshold:

$$log\ p(x_j|M_{\hat{s}}) - log\ p(x_j|M_s) < T_1 \quad (2)$$

In the case of $log\ p(x_j|M_{\hat{s}}) - log\ p(x_j|M_s) = 0$, it means that x_j is non-overlapped feature vectors because it is correctly recognized. Among the feature vectors which were classified into the overlapped category by previous method, if $log\ p(x_j|M_s) - log\ p(x_j|M_{\hat{s}}) < T$, x_j should go to the non-overlapped category, because it might be a feature vector laid on the overlapped region caused by acoustically similar feature between speakers.

Although x_j doesn't satisfy equation (2), it may be a meaningful feature vectors when it satisfies following equation:

$$P_s - P_{avg} > T_2, \quad (3)$$

$$where \quad P_{avg} = \sum_{i=1}^{S} log\ p(x_j|M_i)/S, \ P_s = log\ p(x_j|M_s)$$

It is because that if x_j is a common overlapped feature vector, average likelihood will be almost same to likelihood of correct speaker model. While determining the threshold T_2, we found that T_2 should be changed along with the probability P_s. P_s can be regarded as a distance from the center of a speaker model M_s to x_j. When P_s increases, it means that x_j gets closer to M_s. In this case, although difference between P_s and P_{avg} is small, x_j should be classified into non-overlapped category. When P_s decreases, it means that x_j gets far from M_s. In this case, regardless of the difference between P_s and P_{avg}, x_j should be classified into overlapped category. To change the threshold T_2 related with the P_s, we divided right side of the equation (3) by P_s. As a result, equation (4) is derived as follow:

$$log\ p(x_j|M_s) * (log\ p(x_j|M_s) - P_{avg}) > T_2 \quad (4)$$

Based on our hypothesis, we classify feature vectors from each speaker's training data into two categories as follows:

- x_j : jth input vector, $j = 1,\ldots,N$.

- $\hat{s}$ = argmax $log\ p(x_j|M_i)$, $i = 1,\ldots,S, j = 1,\ldots,N$.
- If $log\ p(x_j|M_{\hat{s}}) - log\ p(x_j|M_s) < T_1$, $x_j \rightarrow P$ (a vector set of a non-overlap category) , where s is a correct speaker index, $T_1$is a threshold.
- Else if $log\ p(x_j|M_s) * (log\ p(x_j|M_s) - P_{avg}) > T_2$, $x_j \rightarrow P$ where $P_{avg} = \sum_{i=1}^{S} log\ p(x_j|M_i)/S$.
- Else $x_j \rightarrow Q$ (a vector set of an overlap category).

After feature vector classification, we reconstruct two speaker models (overlapped and non-overlapped) for each speaker. The test phase is exactly same as system in 3.1.

By using our method, we can use more feature vectors, which can contribute to discrimination power, than baseline system. And we still discard confusing features at the same time. Hence, proposed approach can have better performance.

4. Experiments and results

We performed experiments on a 544-speaker data subset (314 females and 230 males) obtained from the Speaker Recognition Benchmark NIST Speech (1999) corpus. We made three different test cases (8 speakers, 16 speakers, 24 speakers) to show the performance changes related with the number of enrolled speakers. We made 50 test sets for each test case. We used 50s spontaneous speech for each speaker. The front 40s was used for training speaker models and later 10s for testing. Gaussian mixture models (with 16 mixtures) were used for speaker modeling. We extracted 24 dimensional MFCCs from the 8000 Hz sampled signal. We used 30 ms hamming window which was shifted by 10ms.

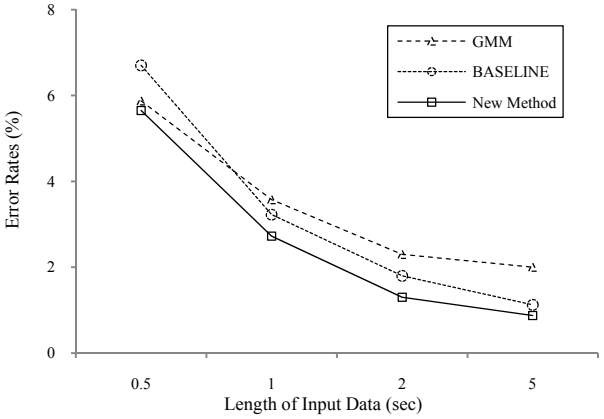
It is known that utterances should be longer than 2s to achieve adequate accuracy in speaker identification [5]. To study the error rates changes related with the test utterances length, we performed experiments on various lengths of speech data (0.5, 1, 2, and 5s spontaneous utterances). To compare with the existing method, we conducted same experiments for three speaker-identification methods (GMM, Baseline system and our proposed method). GMM is the speaker identification system based on the conventional GMM method and baseline system, which is described in section 3.1, means the system based on selective use of feature vectors. For each case, 10s test utterances were divided into short utterances for testing. For example, to identify 8 speakers with 1s utterances, 8 ten-second utterances from 8 speakers were chopped into 80 one-second utterances. In the proposed system, we used threshold $T_1 = 2.0$ and $T_2 = -180.0$ which was found empirically. The error rate was calculated as follows:

$$\text{Error rate} = F_u / T_u, \quad (5)$$

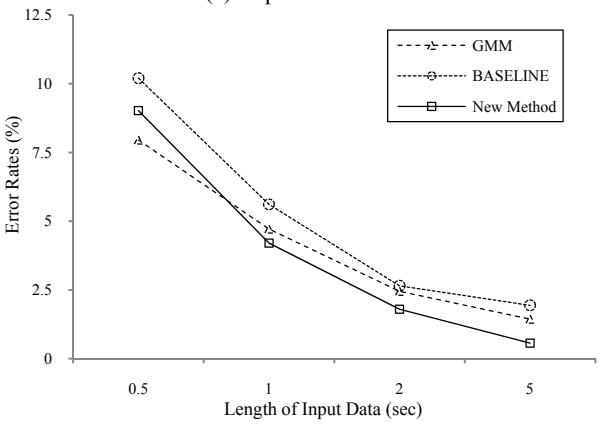
where F_u is the number of falsely identified utterances, and T_u is the total number of utterances.

Fig. 2 shows the experimental error rates for speaker identification test. We computed error rate of each input utterance length with different number of speakers. It is observed that the proposed method outperforms the two methods for all utterance length in 8 speaker test. In 16 and 24 speaker tests, our method shows better accuracy in all utterance length than two methods except 0.5s test. In 0.5s utterance identification, we can get only a limited number of feature vectors after classification. Hence, the accuracy is lowered than conventional GMM method due to the lack of feature vectors.

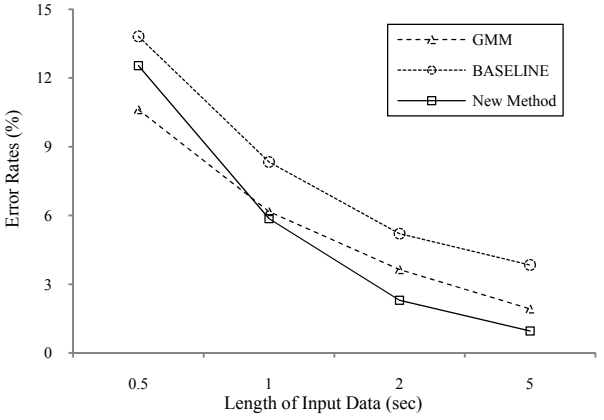
The baseline system shows better performance than conventional method for all utterance length except 0.5s test in 8 speaker test. But in the 16 speaker and 24 speaker tests, error rates were higher than GMM for all utterance length. It's because that the overlapped region caused by acoustic similarity between speakers has been bigger.



(a) 8 speaker test



(b) 16 speaker test



(c) 24 speaker test

Figure 2: Error rates with different number of speakers

Because our method shows better performance than two other methods for longer than 1s tests, it is shown that the proposed method does well in selecting meaningful feature vectors.

There are several issues to be researched. First, results for 0.5s utterances tests with 16 and 24 speakers tell that our method still dumps meaningful feature vectors. Hence, we need to find additional method to improve the feature vector classification performance. Second, we used threshold T_1 as a fixed value. If we can change the threshold automatically by utilizing actual log-likelihood value, number of speakers, and other given information, the discrimination power of the system can be improved. We will perform experiment after changing equation (2) like equation (4). Third, we should do experiments with heavily overlapped data. The baseline system

was known to have better performance than GMM when the speaker models were heavily overlapped [4]. Hence, to prove our method works well when the speaker models are heavily overlapped, we will do experiments with data which contains more background silence or environment noise.

5. Conclusions

The purpose of speaker identification is to select a person when the test utterance is given. Conventional speaker identification system constructs one speaker model for each speaker. In the test phase, overlap regions of speaker models lower the system accuracy.

Recently, a feature vector classification method was proposed to overcome this problem. This method classifies feature vectors from each speaker's training data into two categories: overlapped and non-overlapped, and the system constructs two speaker models for each speaker. In testing phase, the system uses only non-overlapped feature vectors and can get fairly high accuracy. However, this approach has a weak point that the system performance is lower than conventional GMM method when the number of enrolled speaker increases. We thought that this phenomenon was happened because they discard overlapped feature vectors regardless of overlap source.

In spontaneous speech processing, there are three major overlap sources: acoustically similar features of speakers, background silence and environment noise. We made a hypothesis that overlapped feature vectors caused by the acoustic similarity will contribute to make a right decision. To distinguish useful overlapped feature vectors, we proposed a method which classifies feature vectors using log-likelihood ratio in training phase.

Experimental results showed that the proposed method has better performance than baseline system for every utterance tests. Compared with conventional method, the new method showed better performance except 0.5s test. It means that we can use more meaningful feature vectors and discard confusing ones by using our method. In speaker indexing application, overlapped feature vectors caused by common features are the main reason which lowers the system performance. Hence, because our new method does well in classifying feature vectors, it can be used to solve the problem. Now, our ongoing research is focused on expending the proposed method to more general case.

6. References

[1] J.P. Campbell, "Speaker Recognition: A Tutorial", *Proceedings of the IEEE.*, 85(9):1437-1462, 1997.

[2] Reynolds, D.A. and Rose, R.C., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Trans. Speech and Audio Processing.,* 3(1):72-83, 1995.

[3] Bimbot, F., Bonastre, J.F. and et al., "A Tutorial on Text-Independent Speaker Verification", *EURASIP Journal on Applied Signal Processing.*, 4:430-451, 2004.

[4] Kwon, S. and Narayanan, S. "Robust speaker identification based on selective use of feature vectors", *Pattern Recognition Letters.*, 28:85-89, 2007.

[5] Reynolds, D.A., Torres-Carrasquillo, P., "Approach and applications of audio diarization", *In: Proc. IEEE International Conf. on Acoustics, Speech, and Signal Processing.,* , 953–956, 2005.