화자 인식을 위한 특징 벡터의 유연한 선택*

윤 상 민, 박 경 미, 김 길 연, 오 영 환
한국과학기술원 전자전산학과 전산학 전공

Flexible selection of feature vectors for speaker identification

Sangmin Yoon
Korea Advanced Institute of Science and Technology
E-mail : {sangmin, kmpark, kykim, yhoh} @speech.kaist.ac.kr

Abstract

This paper proposes a flexible selection method of feature vectors for speaker identification. In speaker identification, overlapped region between speaker models lowers the accuracy. Recently, a method was proposed which discards overlapped feature vectors without regard to the source causing the overlap.

We suggest a new method using both overlapped features among speakers and non-overlapped features to mitigate the overlap effects.

I. Introduction

In biometric recognition systems, voice is one of the powerful measurements now. There are two main reasons that make voice a better biometric. First, speech is a natural signal to produce that is not considered threatening by users to provide. Second, the telephone system provides a ubiquitous network for obtaining and delivering the speech signal. In addition, amount of spoken documents has been increased and the rate of increase will be faster than before. The importance of speaker identification techniques like speaker indexing, speaker tracking is growing related with this flow.

The task of conventional speaker identification is choosing the most likely person, who was enrolled before, when the test utterance was given. In text-independent case, the most popular speaker models are based on Gaussian mixture models (GMMs) of speech spectral feature [1]. Mel-Frequency Cepstral Coefficients(MFCC) is the most frequently used one.

In general, the accuracy of speaker identification system is related with the amount of given data for both training and testing. When the model is constructed, if the length of the test utterance is long enough, typically 2s or more, the conventional system based on GMMs shows high performance.

However, there are many short speech segments in spoken documents which cause an higher error rates. An Audio data from telephone conversation, meeting often contains short utterances like "Yes", "No" which typically last about 0.5s or fewer. Because short utterances have limited number of feature vectors, data amount is insufficient to make a right decision and easy to affected by specific vectors that are apt to induce decision errors.

In this paper, we suggest a flexible selection method of feature vectors for speaker identification.

Section 2 explains the conventional speaker

* 본 연구는 방위사업청과 국과연의 지원을 받아 2007년도 소프트웨어 설계 특화센터를 통해 수행되었음

identification method and section 3 explains the selective use of feature vectors. Section 4 describes our new method and section 5 shows the experiments and discusses results. Conclusions and future plans are described in section 5.

II. Conventional speaker identification

To verify the identity of a claimed speaker, consider that H0 be the hypothesis that the user is an impostor and let H1 be the hypothesis that the user is the claimed speaker. The scores of the observations are assumed to be generated by random variables characterized by distinct probability density functions according to whether the user is the claimed speaker or an impostor [2].

Let $p(z|H_0)$ be the conditional density function of observation score, z, generated by imposters, and let $p(z|H_1)$ be for the claimed speaker. Then the likelihood ratio is

$$\lambda(z) \equiv p(z|H_0)/p(z|H_1). \quad (1)$$

If $\lambda(z) \geq T$, the decision rule is to choose H_0, otherwise H_1. The threshold, T, is set for a minimum error performance.

In the more general case of identifying a speaker from among N speakers, the decision rule is that speaker i is chosen such that $p_i(z) > p_j(z)$,

$$j = 1, 2, \ldots, N, \; j \neq i, \quad (2)$$

where $p_i(z)$ is the probability of speaker i on input data, z. The speaker with the minimum of error probability is chosen [3].

In speaker recognition system, decision errors are caused by overlapped region between speaker models. When the number of speakers increases, the regions of model overlaps usually increase. This causes to higher error rates. Hence, mitigating overlap effects is important to reduce error rates. Background silence, environment noise and acoustically similar features of speakers are known to cause of speaker model's overlap.

III. Flexible selection of feature vectors for speaker identification

1. Speaker identification based on selective use of feature vectors

Recently, there was a research on a method for robust speaker identification by using selective use of feature vectors [4]. In the paper, to reduce the errors due to overlap, they designed speaker models in a modified way compared with the conventional method. In training phase, they classified feature vectors into two classes for each speaker, non-overlapped and overlapped without regard to overlap source. Then construction of two speaker models for each speaker is performed.

In testing phase, when the speech segment is given, the system extract feature vectors and classified into two classes, non-overlapped and overlapped, by using reconstructed model. Then, the decision was made by maximum likelihood calculation with non-overlapped vectors from non-overlapped speaker models.

2. Flexible selection of feature vectors

Our new idea is focused on feature vector classification step. The major problem of baseline system is that classifying feature vectors without regarding to overlap source. If acoustic feature between speakers becomes more similar the overlap region between speaker models gets bigger. In this case, performance of base line system is uncertainty.

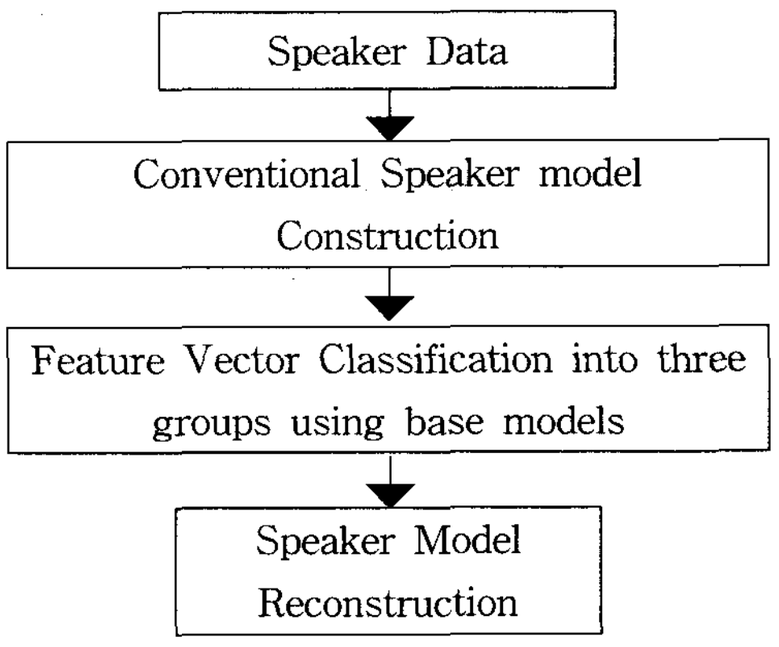

Fig. 1. Block Diagram for speaker modeling training

To build a speaker model more robustly, We classify non-overlapped features into two groups, common-overlapped group(COG) and speaker-similarity group(SSG). COG contains feature vectors

which were caused by environment noise or background silence. SSG contains the feature vectors which were caused by acoustically similar features of speakers. Then SSG is classified into several models. Fig. 1 illustrates the procedure for splitting speaker model.

The procedure is described as follows :

- x_j : j_{th} input vector for training, j=1,...N
- $\hat{i}$ = argmax$Pr(x_j|M_i)$, i=1,...,S, j=1,...,N
  (Mi :conventional speaker model)
- If $\hat{i}$ is a correct speaker index, $x_j \rightarrow P$
  (P : a vector set of a non-overlap category).
- Else if $\hat{k}$ is a correct speaker index and has second maximum $Pr(x_j|M_i)$ then $x_j \rightarrow S_{ik}$
  (S_{ik} : a overlapped vector set which is caused by speaker i,k)
- Else $x_j \rightarrow O$ (O: a vector set of a overlap category)

After feature vector categorization, we reconstruct the speaker models. For each speaker i, we build non-overlapped speaker model(M_i^p), with the vectors of P. By using vectors in S_{ik}, we build overlapped speaker model (M_{ik}) between speaker i and k. Finally, we build an common overlap model M_{co}, with the vectors of O.

Using reconstructed speaker models, we select feature vectors for testing as follows :

- x_j: j_{th} input vector for testing, j=1,...N
- If $\Pr(x_j|M_{co}) > \max Pr(x_j|M_i^p)$ and $\Pr(x_j|M_{co}) > \max Pr(x_j|M_{ik})$, discard feature vector. i,k=1,...,S.
- If $\max Pr(x_j|M_i^p) > Pr(x_j|M_{co})$ and $\max Pr(x_j|M_i^p) > \max Pr(x_j|M_{ik})$, $x_j \rightarrow T_n$, i,k=1,...,S. where T_n is a set of non-overlapped feature vector for testing.
- else $x_j \rightarrow T_s$, where T_s is a set of overlapped feature vector for testing.

When selection is finished, testing procedure occurs as described.

- If $N_{no} \geq N_{so}$, then $\hat{i}$ = argmax $\Pr(x_j|M_i^p)$ is selected. i = 1,...,S
- else $\hat{j}$, $\hat{k}$ is computed by argmax$Pr(x_j|M_{ik})$, $\hat{j}$=j, $\hat{k}$=k. If $\hat{i}$ = argmax $\Pr(x_j|M_i^p)$ is either $\hat{j}$ or $\hat{k}$ then $\hat{i}$ is selected. i,j,k = 1,...S. If $\hat{i}$ is neither $\hat{j}$ or $\hat{k}$ than $\hat{i}$ = argmax($\Pr(x_j|M_{\hat{j}}^p)$,$\Pr(x_j|M_{\hat{k}}^p)$).

Ⅳ. Experiments and results

We performed experiments on a 80-speaker data subset obtained from the STONHENG in The Road Rally Word-Spotting Corpora(RDRALLY1) from NIST(1991). 8 speakers (4 males, 4 females) were randomly chosen from the 80 speakers. we used 50s of spontaneous speech for each speaker: 40s of which were used for training speaker models and 10s for testing. For speaker modeling, Gaussian mixture models (with 16 mixtures) were used. We extracted 24 dimensional Mel-Frequency Cepstral Coefficients from the 10000Hz sampled signal. We used a 30ms Hamming window that was shifted by 10ms.

Typically, we need utterances longer than 2s to achieve adequate accuracy in speaker identification [1]. Hence, we conducted experiments with three speaker-identification methods(conventional GMM, Speaker-identification based on robust feature selection, and our new method). We used various lengths of speech data (0.25s, 0.5s, 1, and 2-s spontaneous utterances).

The error rate was calculated as follows :

$$\text{Error rate} = F_u / T_u, \quad (3)$$

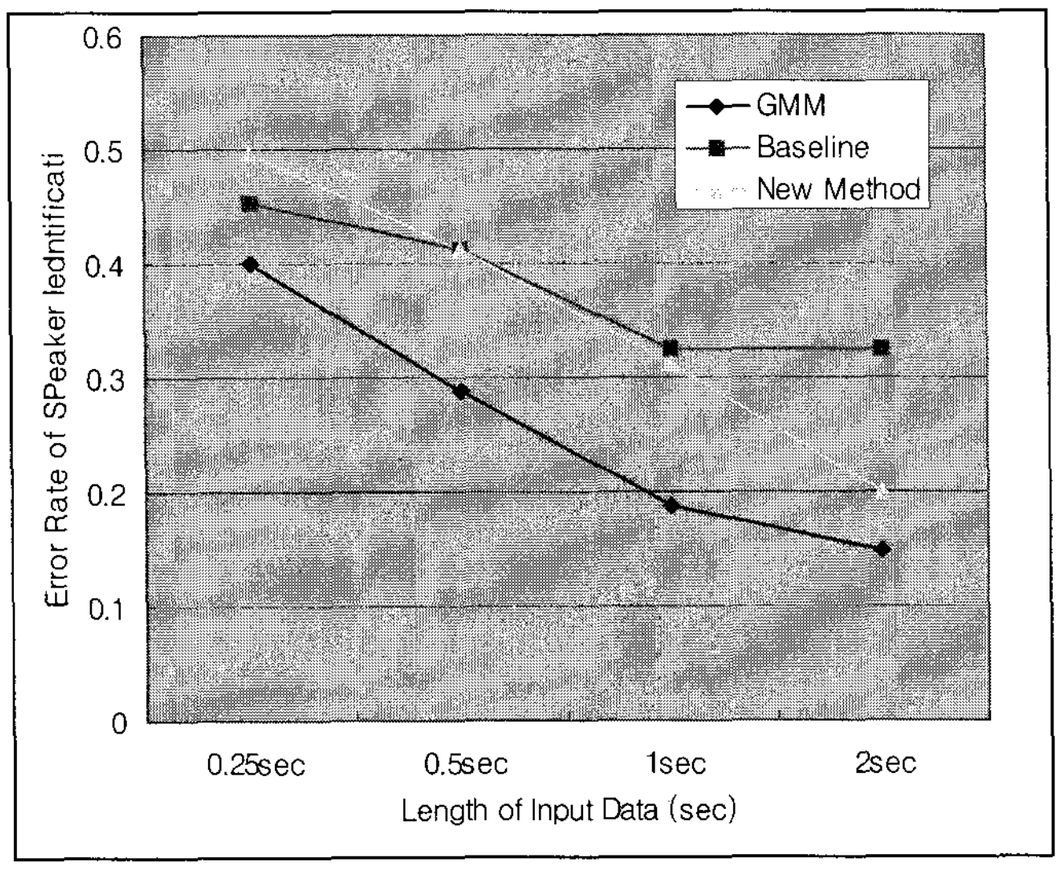

Fig. 2. Error rates for short-segment speaker identification

Fig.2 shows the error rate as a function of input

utterance length. It is observed that GMM method outperforms both the robust speaker identification based on selective use of feature vectors and the proposed method.

In general case, baseline may have a better performance than GMM system[5]. However, when more utterances are classified into overlapped model, overall system performance goes down. The purpose of baseline is selecting robust feature vectors for robust identification. But when most of the feature vectors is classified into overlapped model, due to lack of feature vectors, error rate goes up.

Our purpose is reuse the overlapped feature vectors to maintain performance in various condition. Fig. 2 shows that our method shows better performance than base line at 1 second and 2 second. In 0.25 second test, our method show higher error rate. It's because we just have 40 feature vector set for 0.25 utterance identification. Although we reuse the overlapped feature vectors caused by speaker similarity, we classify them into many speaker model. Hence, the amount of data is too small to make a good decision.

V. Conclusion

The purpose of speaker identification system is extracting speaker information from a sequence of spoken words. When input utterances are long, accuracy of conventional system based on GMMs can be fairly high. But in short speech segments identification, it is easy to be affected by overlap effects which lower accuracy. One method was presented which deals the overlap problem between speakers. Ironically, if the overlap region get's bigger than certain degree the performance goes down.

We described a method which uses feature vectors in flexible way in both training phase and testing phase. To overcome decision errors that arise due to model overlap, we classified feature vectors in three groups. And we used robust features, some useful overlapped features when they are needed. For useless features, we discarded them.

To achieve a high quality speaker identification several related works may followed. First, we should figure out when the GMMs works well or not compare to baseline. Second, we should make an robust speaker identification modeling technique by using additional method or new method. Finally, we should apply our method in longer data identification.

Reference

[1] D.A. Reynolds, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Trans. Speech Audio Process. 3 (1), 72-83, 1995.

[2] Rosenberg, A.E., Siohan, O., Parathasarathy, S., "Speaker verification using minimum verification error training." Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, 105-108 In Proc

[3] J.P. Campbell, "Speaker Recognition: A Tutorial," Proc. IEEE 85, 1437-1462, 1997.

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

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