Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition Letters 28 (2007) 85–89

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Robust speaker identification based on selective use of feature vectors

Soonil Kwon [a,*], Shrikanth Narayanan [b]

[a] *System Technology Division, Korea Institute of Science and Technology, Seoul 130-650, Korea*

[b] *Department of Electrical Engineering and IMSC, University of Southern California, Los Angeles, CA 90089, United States*

Received 7 June 2005; received in revised form 5 June 2006
Available online 21 August 2006

Communicated by O. Siohan

Abstract

A new method for speaker identification that selectively uses feature vectors for robust decision-making is described. Experimental results, with short speech segments ranging from 0.25 to 2 s, showed that our method consistently outperforms other approaches yielding relative improvements of 20–51% and 15–30% over baseline GMM and the LDA-GMM systems, respectively.

Keywords: Short-segment speaker identification; Speaker model construction; Feature vector selection; Linear discriminant analysis (LDA)

1. Introduction

In speaker identification, the conventional decision rule is to choose the speaker with the minimum probability of error. The most popular speaker models are based on Gaussian mixture models (GMMs) of speech spectral features, typically, Mel-Frequency Cepstral Coefficients. The performance, among other factors, depends on the amount of data available for identification. In general, experimental evidences have shown that utterances need to be long enough to ensure capturing adequate speaker discrimination. If the spoken utterance is adequately long, typically 2 s or more, state of the art methods such as those based on GMMs achieve fairly high accuracy. Although speaker models tend to be overlapped, if given enough data, it is not likely to have a serious effect on speaker identification using the conventional method. However, in some applications, it is desirable, or necessary, to operate with short speech segments. For example, in speaker indexing scenarios for speech data monitoring and mining (Yang et al., 1999; Kwon and Narayanan, 2005), we sequentially process data to verify who is talking while using short segments of speech without missing any speaker changes within that segment.

For instance, consider audio data from spontaneous speech interactions such as telephone conversations or meetings. Such data often contain short utterances from some of the speakers, representing things such as acknowledgement and back channel events (e.g., "Yes", "No", or "Sure") which typically last about 0.5 s or fewer. In general in such spontaneous dialogs, we may have to identify many such segments lasting fewer than 2–5 s, contributing to higher speaker identification error rates (Reynolds and Carrasquillo, 2005). There are two issues we face under this circumstance. The first is due to insufficient data to afford acceptable speaker discrimination. The second relates to robustness in the sense that a smaller data set is usually more susceptible to specific feature vectors that are apt to induce decision errors. In this paper, we assume that we may have to deal with short data segments and address the problem of robust speaker identification. We propose a simple method for creating speaker models wherein we attempt to eliminate feature vectors that can potentially cause identification errors.

We performed some experiments to evaluate our method. Usually there are a varying number of people

[*] Corresponding author. Tel.: +82 2 958 5606; fax: +1 213 740 4651.
E-mail address: soonilkw@yahoo.com (S. Kwon).

0167-8655/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2006.06.009

participating in conversations such as meetings and debates. In our experiments, we assume the maximum number of participants is 8. We created 50 different data sets. Each set consists of 8 speakers, 4 males and 4 females. We identified these 8 speakers using relatively short utterances: 0.25, 0.5, 1, 2 s. Each utterance represents spontaneous speech from telephone conversations. Experimental results showed that our new method consistently achieves higher accuracy than the conventional method.

The rest of this paper is organized as follows: Section 2 explains the conventional speaker-identification method; Section 3 describes our new method; Section 4 describes the experiments and discusses results; conclusions and future plans are described in Section 5.

2. Conventional speaker identification

Speaker recognition problems are typically formulated as hypothesis testing. For example, to verify the identity of a claimed speaker, consider that H_0 be the hypothesis that the user is an impostor and let H_1 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 (Rosenberg et al., 1998).

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

$$\lambda(z) \equiv p(z|H_0)/p(z|H_1). \tag{1}$$

If $\lambda(z) \geqslant T$, 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

$$\text{speaker } i \text{ is chosen such that } p_i(z) > p_j(z),\quad j = 1, 2, \ldots, N,\ \ j \neq i, \tag{2}$$

where $p_i(z)$ is the probability of speaker i on input data, z. The speaker with the minimum of error probability is chosen (Campbell, 1997).

In either the verification or the more general identification scenario, the decision critically depends on the interrelation between the competing probability density functions. Specially, overlap regions of probability density functions (or speaker models) contribute to decision errors. Usually, with increasing number of speakers, the regions of potential model overlap increase, resulting in more decision errors. Hence it becomes important to reduce the effect of these overlapped regions.

3. Speaker identification based on the selective use of feature vectors

In spontaneous speech processing, the overlap of speaker models is usually caused by background silence, environment noise, and acoustically similar features of speakers. For instance, silence and environment noise, which might be common features in a data stream of a session, would hence be present in the feature space of all of the speakers to be identified. The influence of such common features likely to contribute to model overlap is especially critical when the utterance length (data available for identification) is limited. In this paper, we focus on mitigating the overlap effects, without regard to the source causing the overlap, for robust speaker identification with short utterance lengths.

Fig. 1 illustrates the difference between conventional speaker models and the new speaker models with reduced overlap. The vectors on the left side of B are always recognized as those from Speaker-1 (original) while the vectors on the right side of B are always recognized as those from Speaker-2 (original). The decision-making like this has unavoidable errors. As the testing utterances get shorter, it becomes more vulnerable to identification errors. The idea here is to select relatively robust feature vectors and use them for decision-making.

To reduce the errors due to overlap, we design speaker models in a modified way compared with the conventional method. We split each speaker model into two models: non-overlapped and overlapped (Fig. 1). Fig. 2 illustrates the procedure for splitting speaker models for the case of identifying 2 speakers. Firstly, we train using the standard approach two speaker models (GMMs) with two speaker-specific speech data sets. Next, using the maximum likelihood method with the base speaker models built in the previous step, each feature vector is verified if it can be correctly classified. There could be some vectors falsely recognized if competing speaker models are overlapped. We classify the feature vectors from each speaker into two categories: non-overlap and overlap. In the last step of training, based on the reclassified feature vectors, we reconstruct two models for each speaker: non-overlapped and overlapped speaker models.

Fig. 1 shows two models for each speaker in solid line. The feature vectors included in the range between A and

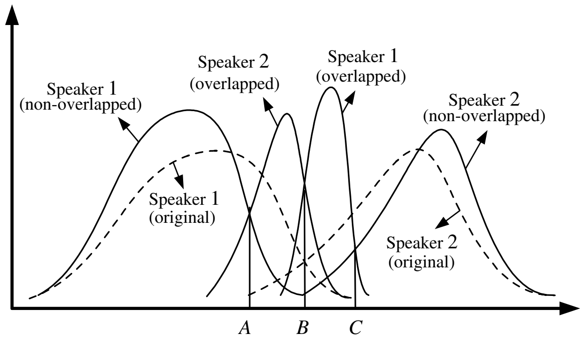

Fig. 1. Hypothesis testing for two speaker identification with assumption that speakers can be represented by one-dimensional statistical models: dashed line for conventional speaker models and solid line for the proposed speaker models based on the selective use of feature vectors.

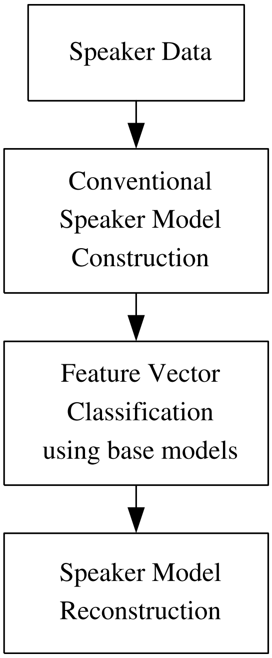
Speaker Data
Conventional Speaker Model Construction
Feature Vector Classification using base models
Speaker Model Reconstruction

Fig. 2. Block diagram for speaker model training.

C are ignored for decision-making (Fig. 1). From the input utterances, we pick out feature vectors only from the left side of *A* and the right side of *C* and use them for decision-making (Fig. 3).

For example, assume there are *S* single-speaker speech-data sets. With feature vectors extracted from these data, we train speaker models, M_i, where $i = 1,\ldots,S$. Then we categorize feature vectors from each speaker data into non-overlapped and overlapped vectors using a maximum likelihood criterion as follows:

- x_j: *j*th input vector, $j = 1,\ldots,N$.
- $\hat{i}_j = \arg\max Pr(x_j|M_i)$, $i = 1,\ldots,S$, $j = 1,\ldots,N$.
- If $\hat{i}_j$ is a correct speaker index, $x_j \to P$ (*a vector set of a non-overlap category*).
- Else $x_j \to Q$ (*a vector set of a overlap category*).

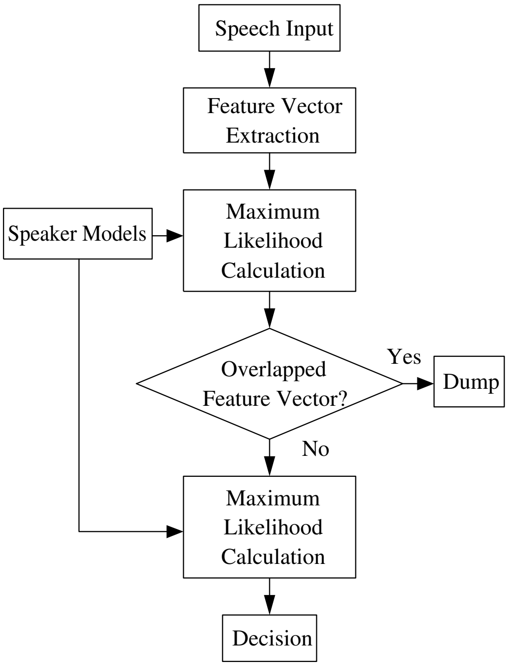
Speech Input
Feature Vector Extraction
Speaker Models
Maximum Likelihood Calculation
Overlapped Feature Vector?
Yes
Dump
No
Maximum Likelihood Calculation
Decision

Fig. 3. Block diagram of speaker identification using the selected feature vectors.

After feature vector categorization, we reconstruct the speaker models. For each speaker *i*, we build two models, non-overlapped (M_i^P) and overlapped (M_i^Q), with the vectors of *P* and the vectors of *Q*, respectively. Using the pairs of speaker models, we select feature vectors for testing as follows:

- x_j: *j*th input vector for testing, $j = 1,\ldots,N$.
- If $\max Pr(x_j|M_i^P) > \max Pr(x_j|M_i^Q)$, $i = 1,\ldots,S$. $x_j \to T$, where *T* is a set of selected feature vectors for testing.
- Check all of x_j.
- $\hat{i} = \arg\max Pr(T|M_i^P)$, $i = 1,\ldots,S$.

Finally, $\hat{i}$ is the index of a speaker identified with only non-overlapped speaker models.

Our new method can be very useful for sequentially identifying speakers with short utterances. Some feature vectors may lie where there is overlap with other speaker models. When those vectors are included within the short utterance, it is likely to contribute toward a wrong decision. By splitting speaker models, we can select feature vectors to reduce the risk of decision errors.

4. Experiments and results

We performed experiments on a 400-speaker data subset (200 females and 200 males) obtained from the Speaker Recognition Benchmark NIST Speech (1999) corpus. We made 50 sets consisting of 8 speakers (4 males and 4 females) randomly chosen from the 400 speakers. There are 50 s of spontaneous speech for each speaker: 40 s of which were used for training speaker models and 10 s for testing. For speaker modeling, Gaussian mixture models (with 16 mixtures) were used. We extracted 24 dimensional Mel-Frequency Cepstral Coefficients from the 8000 Hz sampled signal. We used a 30 ms Hamming window that was shifted by 10 ms.

Typically, we need utterances longer than 2 s to achieve adequate accuracy in speaker identification (Reynolds and Rose, 1995). Hence, to study the relation between the error rate and the length of short utterances, we conducted experiments with three speaker-identification methods (conventional GMM, GMM with LDA, and our new method) on various lengths of speech data (0.25, 0.5, 1, and 2-s spontaneous utterances). For each case, 10-s testing utterances to be identified were chopped into short utterances for testing. For example, to identify 8 speakers with 1-s utterances, 8 ten-second utterances from 8 speakers

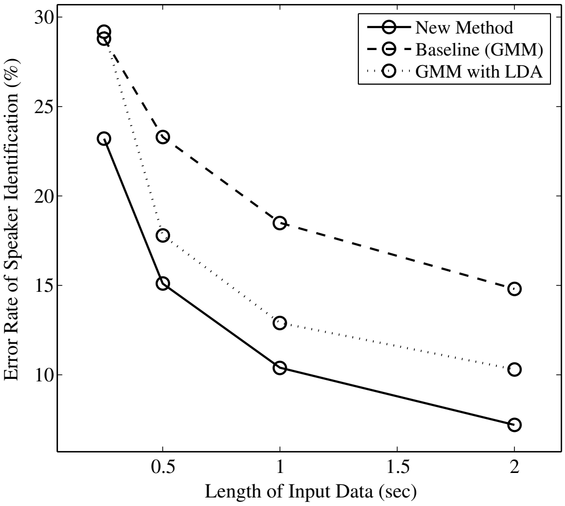

Fig. 4. Error rates for short-segment speaker identification.

were chopped into 80 one-second utterances. The error rate was calculated as follows:

$$\text{Error rate} = F_u/T_u, \tag{3}$$

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

Fig. 4 shows the experimental speaker identification error rate as a function of input utterance length. It is interesting to observe that the new method outperforms the conventional GMM method for all the utterance lengths considered. The range of the difference in error rate between the GMM (baseline) and our new method is from 5.6% to 8.2% absolute with various lengths (from 20% to 51% relative). The error rate of our method with 0.5 s utterances (15.1%) is almost the same as the error rate of the conventional method (baseline) with 2 s utterances (14.8%). It means that we can achieve about 85% accuracy with one fourth the length of utterances using the new method. An error rate improvement of 5.6% absolute (20% relative) was obtained for the shortest utterance length considered (0.25 s). In addition, it is remarkable that our method outperforms another conventional approach, GMM with LDA, even though LDA is a dimension-reducing method to maximize discriminating powers (Jin and Waibel, 2000). The improvement in error rate between the GMM with LDA and our new method is from 2.5% to 6.0% absolute (from 15% to 30% relative). These results experimentally confirm that the new method can yield robust speaker identification with short data segments.

There are several issues that remain to be investigated. First, an analysis that relates identification performance as a function of degree of overlap needs to be done. Certain common features may not have as negative an effect if the competing speaker models are not much overlapped. In this case, both the conventional and the new method are expected to do well. On the other hand, if competing speaker models are heavily overlapped, both methods are expected to degrade. Table 1 shows an experimental result that the ratio of overlapping and non-overlapping feature vectors (ROV) goes higher as the length of segment gets shorter. In the cases of 0.5 and 0.25 segments, ROV could be 1 that means every vector within a segment was classified into the overlap category: We considered those cases as identification errors. However, it is a rare occurrence (up to 2.5%). We need to systematically characterize the performance bounds as a function of overlap for various data lengths.

Table 1
The relation between the ratio of overlapping and non-overlapping feature vectors (ROV) and the length of segment: ROV is the number of overlapping vectors divided by the total number of vectors

Ratio of overlap (ROV)	Length of segment (s)			
	0.25	0.5	1	2
$0.5 \leqslant \text{ROV} < 0.7$	23.3%	25.4%	30.1%	31.6%
$0.7 \leqslant \text{ROV} < 0.9$	12.0%	13.1%	7.7%	4.2%
$0.9 \leqslant \text{ROV} < 1$	5.1%	1.7%	0.5%	0.3%
$\text{ROV} = 1$	2.5%	0.06%	0%	0%

Second, we considered the scenario of identifying only 8 speakers in the experiments of this paper with the assumption that the number of participants of conversation or meetings usually smaller than eight. However, as the number of speakers increases, the overlapped regions also increase. In this case, the number of robust vectors available for identification from a given data segment decreases. Hence, we need to investigate the performance bounds as the number of speakers increases. It should be noted that this problem depends on the characteristics of competing speakers and environmental conditions. Finally, we need to investigate the performance of our method with respect to various signal conditions (including different Signal-to-Noise Ratios). These are topics of our ongoing work.

5. Conclusions

The speaker identification process aims at extracting speaker information from a sequence of spoken words. The identification performance needs to contend with issues arising due to overlap amongst the models owing to a number of factors including common acoustic environment and speaker similarities. When input utterances are long (typically longer than 5 s), speaker identification accuracy can be fairly high. However, we frequently have problems in identifying speakers from spontaneous speech data such as from telephone conversations and meetings where some of the speakers' utterances can be very short (fewer than 2 s). One problem with short segments is that the performance is especially vulnerable to feature vectors likely to cause errors due to model overlap.

We described a simple method that employs only feature vectors that are deemed to contribute to discrimination. To overcome decision errors that arise due to model overlap, we trained speaker models to separate the data and select only useful feature vectors for more accurate speaker identification. Experimental results showed that this approach helped improve the speaker identification performance in overcoming some of the difficulties arising when speaker models appear overlapped in a given feature space. We also showed that the method outperformed the conventional approaches for all data lengths considered including short utterances. The method is hence useful for detecting speakers from short segments in speech indexing applications as well as for improved performance for rapid speaker identification. Additionally, the method promises superior performance for longer data segments as well, and future experiments will focus on further validating this aspect of the results.

References

Campbell, J.P., 1997. Speaker recognition: A tutorial. Proc. IEEE 85, 1437–1462.

Jin, Q, Waibel, A., 2000. Application of LDA to speaker recognition. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP-2000), vol. 5, pp. 250–253.

Kwon, S., Narayanan, S., 2005. Unsupervised speaker indexing using generic models. IEEE Trans. Speech Audio Process. 13 (5), 1004–1013, Part 2.

Reynolds, D.A., Torres-Carrasquillo, P., 2005. Approach and applications of audio diarization. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, V, pp. 953–956.

Reynolds, D.A., Rose, R.C., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. Proc. IEEE Trans. Speech Audio Process. 3 (1).

Rosenberg, A.E., Siohan, O., Parathasarathy, S., 1998. Speaker verification using minimum verification error training. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, pp. 105–108.

Yang, J., Zhu, X., Gross, R., Kominek, J., Pan, Y., Waibel, A., 1999. Multimodal people ID for a multimedia meeting browser. In: Proc. 7th ACM Internat. Conf. on Multimedia, Part 1, pp. 159–168.