Ensemble Classifiers Using Unsupervised Data Selection for Speaker Recognition

Chien-Lin Huang\textsuperscript{1}, Chiori Hori\textsuperscript{1}, Hideki Kashioka\textsuperscript{1} and Bin Ma\textsuperscript{2}

\textsuperscript{1}National Institute of Information and Communications Technology, Kyoto, Japan
\textsuperscript{2}Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore
\{chien-lin.huang, chiori.hori, hideki.kashioka\}@nict.go.jp, mabin@i2r.a-star.edu.sg

Abstract
This paper presents an approach with ensemble classifiers using unsupervised data selection for speaker recognition. Ensemble learning is a type of machine learning that applies a combination of several weak learners to achieve an improved performance than a single learner. Based on its acoustic characteristics, the speech utterance is divided into several subsets using unsupervised data selection methods. The ensemble classifiers are then trained with these non-overlapping subsets of speech data to improve the recognition accuracy. Our experiments on the 2008 and 2010 NIST Speaker Recognition Evaluation datasets show that using ensemble classifiers substantially reduces DCF.

Index Terms: speaker recognition, ensemble classifier, unsupervised data selection

1. Introduction
Demand continues to increase for speaker recognition technology in such applications as telephony, security, and communication. The major components of speaker recognition, which finds the identity information of a speaker from speech signals, include feature analysis, statistical modeling and verification decision. Gaussian mixture model (GMM) has been commonly applied for statistical modeling in speaker recognition applications with speaker adaptation techniques based on maximum a posteriori (MAP) \cite{1}, eigenvoice \cite{2}, and eigenchannel \cite{3}.

In the common adopted speaker recognition framework, Mel frequency cepstral coefficients (MFCCs) are used as the feature analysis and the log-likelihood ratio between the target speaker GMM and the universal background model (UBM) is used to verify the speaker. The advantage of the GMM-based approach is that speaker recognition can be performed in a completely text-independent manner \cite{4}. Conventionally, since all speech frames are used to estimate speaker information and build GMMs, we do not consider data segmentation and selection. However, one disadvantage of such a GMM modeling approach is that the acoustic variability of phonetic events is not taken into account during comparisons with different speakers \cite{4}. To solve this problem, many previous studies focused on using specific constrained groups of data to improve the speaker recognition performance.

Sturim et al. presented text-constrained Gaussian mixture models to close the gap between text-dependent and text-independent speaker verification. Speech is segmented into acoustic units such as words or phones, and then GMM-UBM verifiers are trained and tested using only speech from constrained groups of units \cite{5}. Park et al. proposed speaker identification using domain dependent automatic speech recognition to provide phonetic segmentation. A combination of classifiers is used to reduce identification errors on single test utterances \cite{4}. Baker et al. studied GMM modeling using multilingual broad phonetics to construct syllabic events and segmentations for speaker verification \cite{6}. Bocklet et al. described a speaker recognition approach using syllable-level constraints for cepstral frame selection. Complementary information and improvement can be found by combining eight subsystems including syllable onsets, syllable nuclei, syllable codas, syllables following pauses, one-syllable words, and three other kinds of syllables \cite{7}. Recently, Sanchez et al. studied the performances between constraint-dependent and constraint-independent approaches for training UBMs and joint factor analysis. They explored unit-based constraints which are regions constrained by specific syllables, phones, or sub-phone regions \cite{8}.

All of the above work segmented and selected data for more detailed speaker model construction based on prosody, syllable or phoneme analysis. Although these approaches showed improvements in speaker recognition, many shortcomings remain in them. For example, the quality of the feature frame selection is obviously influenced by the accuracy of automatic speech recognition (ASR) or prosody estimation systems. In addition, prior knowledge is required for such constrain-based approaches as language information.

In this study, we propose an ensemble learning using unsupervised data selection which considers acoustic variability in GMM training for speaker recognition. The speech data are segmented into several subsets of speech frames without any auxiliary information or pre-processor (ASR or prosody estimator systems). The frame counts of the subsets are used as the weights for a combination of ensemble classifiers. We conducted our experiments on the 2008 and 2010 NIST Speaker Recognition Evaluation (SRE) datasets in the GMM-UBM framework.

The rest of this paper is organized as follows. Section 2 presents our proposed ensemble learning using unsupervised data selection. The experimental protocol and results are presented in Sections 3 and 4. Finally, Section 5 concludes this work.

2. Ensemble Classifiers Using Unsupervised Data Selection
In daily life, we usually consult several experts before making an important decision. Ensemble-based systems weigh several opinions and combine them to reach a final decision instead of a single-expert system \cite{9}, \cite{10}. Figure 1 illustrates the proposed ensemble classifiers for speaker recognition. The original speech data are segmented into several data subsets from which several weak classifiers are trained by non-overlapping segmentations. We considered two factors for building the ensemble classifiers for speaker recognition. One is to split and select data based on acoustic variability. The other is to effectively combine the results of ensemble classifiers. In this study, we investigate both the data selection and the combination of ensemble classifiers without any
**2.3. Combination of ensemble classifiers**

With conventional GMM-UBM architecture, the speaker recognition decision is based on the log-likelihood ratio (LLR) between target speaker GMM $A_{SPK}$ and UBM $A_{UBM}$

$$\Lambda = \frac{1}{T} \sum_{t=1}^{T} \left[ \log p(x_t | A_{SPK}) - \log p(x_t | A_{UBM}) \right]$$

where $T$ means the total frames. If the score exceeds threshold $\Lambda > \Theta$, the claimed speaker will be accepted, or else rejected. To exploit the ensemble classifiers in the GMM-UBM architecture, the frame count is used to weight the results of individual weak classifiers. The LLR score is then estimated as follows:

$$\Lambda' = \frac{1}{T} \sum_{k=1}^{K} n_k(X) \times \left[ \log p_k(X | A_{SPK}) - \log p_k(X | A_{UBM}) \right]$$

where $n_k(X)$ is the number of frames in classifier $k$ and satisfies $\sum_{k=1}^{K} n_k(X) / T = 1$. In other words, the contribution of ensemble classifier $k$ is zero if frame count $n_k(X)$ is zero.

### 3. Experimental Protocol

The NIST SRE data were collected from such different types of channel as telephones and microphones. We evaluated the system on the core condition of the 2010 NIST SRE in the tel-tel condition (det5) and all the English trials of the SRE-2008 core task.

#### 3.1. Feature analysis

We applied long-term feature analysis (LTF) [11] as the feature extraction based on the traditional MFCCs of a short-time spectral analysis of 16 ms. We extracted 36 MFCCs consisting of 12 coefficients in addition to the first and second derivatives. Speech signals were divided into 18 sub-bands between 250 to 3500 Hz using the Mel-filter bank to make spectral contents that resemble those of telephone channels. LTF is used to average several short-time spectral features in a long-time window and capture the spectral statistics over a long period of time. We applied the overlapping long-term windows on the short-term features and reduced the number of short-term MFCC frames $M$ to the number of LTF frames $N$ with $N = (M - L) / Z + 1$. $L$ denotes the size of the long-term window and $Z$ is the step of the long-term window shift. With long-term feature analysis, a sequence of short-term feature
vector $c_i$ is used to derive long-term feature vector $c_i = \sum_{l=1}^{L} z_{l} x_{l} / L$. Since the mean of multiple short-term spectral features is used, LTF can simultaneously take account of short-term frequency characteristics and long-term resolution. This transformation results in a more compact feature vector for statistical modeling. The optimal values of $L$ and $Z$ were 4 and 2, respectively, based on our previous study [11].

3.2. Statistical modeling and verification

NIST SRE-2004, SRE-2005, and SRE-2006 one-side data were used to train the gender-dependent UBMMs. The iterative EM algorithm was adopted to estimate the parameters of the Gaussian components. Speaker adaptation is applied because a sufficient amount of training data is not always available. In this study, we tested the ensemble systems with maximum a posteriori (MAP) and eigenchannel adaptation technologies [3]. The eigenchannel adaptation considers various channel factors and provides a good solution for channel mismatch. The eigenchannel assumes the means of the speaker’s model is much smaller than the number of imposter samples [8]. The speaker verification results are reported in terms of $100 \times DCF$ and $1000 \times DCF$ for SRE-2008 and for SRE-2010, respectively.

4. Results

We evaluated the robustness of the ensemble classifiers using unsupervised data selection from several viewpoints.

4.1. Unsupervised data selection

To determine the effect of unsupervised data selection and the ensemble classifiers, we first compared partitioning and hierarchical clustering algorithms based on Euclidean and Mahalanobis distances, and weighting schemes using equal weighting and frame counts. The summarized results were shown in Table 1. The mixture number of UBM was 256. Four subsets were used for the ensemble classifiers with eigenchannel and ZT-norm. The baseline was trained and tested with all the data. We found improvements of the ensemble classifiers using the unsupervised data selection in Table 1. The K-means clustering algorithm outperformed the hierarchical method. The best performance was shown on the K-means algorithm with Mahalanobis distance measurement. The combination of ensemble classifiers with a weighting scheme of frame counts is better than equal weighting. As a result, we applied these settings in the following experiments.

4.2. Ensemble of GMM-UBM systems

GMM-UBM with MAP adaptation is a commonly adopted framework for speaker recognition. Without channel compensation and score normalization, we applied our proposed unsupervised data selection for three data sets including the UBM training data, the speaker enrollment, and the testing data. Each of the ensemble systems used the corresponding data subset segmented by clustering. We conducted experiments with five different numbers of data subsets (2, 4, 8, 16 and 32) on four different numbers of UBM mixtures (128, 256, 512, and 1024). Figure 2 showed the EER and DCF curves on NIST SRE-2010. The number of the zero subset means the conventional approach without ensemble classifiers which is the baseline system trained and tested with all the data. From all the systems in Figs. 2(a) and (b), we found that eight subsets achieved the lowest EER results and 16 subsets gave the best $DCF$ results on average. A special case was UBM with 128 mixtures, which had the best performance with 32 subsets. More data subsets could be applied due to the small size of the UBM mixtures. In all the
conducted GMM-UBM and eigenchannel speaker recognition ensemble classifiers based on the selected data subsets and enrollment, and testing. We trained and evaluated the Mahalanobis distance measurement for UBM training, using a hierarchical or a requirement. The speech data were divided into several subsets SRE-2008 answer keys. The experimental results were shown obtained from the SRE-2010 data using version 3 of the NIST the SRE-2008 core task based on the optimized setting. We evaluated the conversational telephone English speech of numbers of subsets increased. Due to data sparsity, a smaller number of subsets in the ensemble should be applied if a larger size of UBM mixtures is adopted. As a result, UBM with 128 mixtures and eight subsets, and UBM with 256 mixtures with four subsets achieved the lowest DCF scores in Fig. 3(a). Figure 3(b) showed the improvements based on the individual DCF reduction (from $DCF=0.5603\%$ of eigenchannel with ZT-norm to $DCF=0.4435\%$) using the ensemble classifiers with the UBMs of 256 mixtures.

4.4. Results of NIST SRE-2008 core task

We evaluated the conversational telephone English speech of the SRE-2008 core task based on the optimized setting obtained from the SRE-2010 data using version 3 of the NIST SRE-2008 answer keys. The experimental results were shown in Table 2. There was a 14.97% DCF reduction (from $DCF=1.7196\%$ of the baseline system of eigenchannel with the UBMs of 512 mixtures to $DCF=1.4622\%$) using the ensemble classifiers on the SRE-2008 data. The experiment confirmed that the ensemble classifiers consistently improved the speaker recognition performance.

5. Conclusions

We studied the ensemble method using unsupervised data selection for effective speaker recognition. Unlike previous constrain approaches, we had no auxiliary information requirement. The speech data were divided into several subsets using a hierarchical or a $K$-means algorithm with Euclidean or Mahalanobis distance measurement for UBM training, eigenchannel estimation, score normalization, speaker enrollment, and testing. We trained and evaluated the ensemble classifiers based on the selected data subsets and combine the ensemble systems on the frame count. We conducted GMM-UBM and eigenchannel speaker recognition on the NIST SRE-2010 and SRE-2008 data sets. Based on our experimental results, the ensemble classifiers with unsupervised data selection are helpful for speaker recognition.

6. References