AN ACOUSTIC SEGMENT MODEL APPROACH TO INCORPORATING TEMPORAL INFORMATION INTO SPEAKER MODELING FOR TEXT-INDEPENDENT SPEAKER RECOGNITION

Yu Tsao [1], Hanwu Sun [2], Haizhou Li [2], and Chin-Hui Lee[3]

[1] National Institute of Information and Communications Technology, Kyoto, Japan
[2] Institute for Infocomm Research, A*Star, Singapore 138632, Singapore
[3] School of ECE, Georgia Institute of Technology, Atlanta, GA, 30332-0250, USA

ABSTRACT

We propose an acoustic segment model (ASM) approach to incorporating temporal information into speaker modeling in text-independent speaker recognition. In training, the proposed framework first estimates a collection of ASM-based universal background models (UBMs). Multiple sets of speaker-specific ASMs are then obtained by adapting the ASM-based UBMs with speaker-specific enrollment data. A novel usage of language models of the ASM units is also proposed to characterize transitions among ASMs. In the testing phase the ASM sets for the claimed speaker and UBMs, along with a bigram ASM language model, are used to calculate detection scores for each given test utterance. We report on speaker recognition experiments using the NIST 2001 SRE database. The results clearly indicate that the proposed ASM-based method achieves a notable improvement over the GMM-based speaker modeling in which no temporal modeling is considered. Moreover, a further error reduction is obtained by integrating the language model, another inclusion of temporal properties made possibly by ASM based speaker modeling.

Index Terms—Speaker recognition, acoustic segment model

1. INTRODUCTION

Speaker recognition includes two categories of operation, speaker identification and speaker verification. The former aims at identifying a particular speaker while the later intends to authenticate a person's claimed identity. Both categories build on theory of pattern classification and verification using statistical models to characterize each speaker's individuality. Among these models, Gaussian mixture models (GMMs) have been widely adopted in the case when no text information about the spoken materials is given. Good performance on many tasks has been reported. For a GMM-based speaker modeling framework, we use a GMM universal background model (UBM). Accordingly, such system is also known as the GMM-UBM system [1, 2].

In recent years, many techniques have been proposed to enhance the conventional GMM-UBM system. One way to exploit long-time acoustic features is to use temporal information in speech signals [3]. In the GMM-UBM framework these temporal features average out during GMM training. Experimental results verified that such features indeed add some benefit to discriminating speakers and thus enhance the GMM-UBM performance [3].

Another way of incorporating temporal information into speaker modeling is to extend frame-based GMMs to segment models. In 1988, an acoustic segment model (ASM) framework was first proposed to characterize fundamental sound units and acoustic lexicons [4] for automatic speech recognition. Conceptually speaking, each ASM is modeled with a hidden Markov model (HMM) [5]. However one important distinction from phone-based HMMs is that ASMs do not need any phonetic definition, and thus they are capable of modeling speakers in a text-independent manner. With ASMs, we can decode an arbitrary utterance into a sequence of acoustic units, with each token representing a class of acoustic sounds [4]. In recent studies, ASMs have been applied to pattern classification and verification applications, such as spoken language recognition [6] and music retrieval [7]. The capability of modeling temporal information without a direct phonetic mapping enables the ASM-based framework to achieve quite good performance.

In this paper, we utilize the above properties of ASMs to model speakers and the universal background. The proposed ASM-UBM framework can be considered as an enhancement over conventional GMM-UBM speaker modeling. Because multiple ASMs are used here we further model temporal information for speakers and background with language models characterizing ASM unit transitions.

2. ACOUSTIC SEGMENT MODELING

ASM training includes two major steps, initialization and model training processes. Given all training utterances, a rough segmentation and clustering is performed in the initialization stage to obtain a preliminary set of ASM unit transcriptions followed by an iterative training algorithm to refine ASM parameters and segment labels [4, 6, 7].

2.1. Initialization Stage

The initialization stage comprises of two key phases: segmentation and quantization. In the segmentation phase, all training data are partitioned into initial acoustic segments. In this study, we simply perform an even segmentation on all training utterances. Since ASM training is an iterative process, we believe such a rough segmentation is satisfactory. Other algorithms, such as maximum likelihood segmentation [8], can also be used.

Next in the segment quantization phase, we group the initial segments into a small number of sound classes to represent the acoustic space through a segment clustering (SC) algorithm [4]. With K training segments $\{Y_1,\ldots Y_K\}$, we intend to find a segment codebook of C centroid vectors, $\{m_{1,\ldots}m_C\}$, by minimizing the accumulated distortion, D:

$$D = \sum_{k=1}^{K} \min_{c\in 1,\ldots C} d(Y_k, m_c), \quad (1)$$

where $d(Y_k,m_c)$ is the segment distortion between codeword m_c and segment Y_k. More details about the segment quantization algorithm can be found in [4].

2.2. Model Training Stage

The model training stage consists of three steps: ASM initialization, decoding, and refinement. With acoustic segments and their corresponding tokens provided by the preceding initialization stage, we can build an initial set of ASMs by grouping segments with the same label. Here we assume each ASM is characterized by an HMM, so we can decode all training data into sequences of ASM units with Viterbi decoding. ASM refinement can now be performed by iteratively re-estimating ASM parameters with the Baum-Welch algorithm [5]. In general three to five iterations of re-estimation is enough to reach convergence.

3. ASM-BASED SPEAKER RECOGNITION

The ASM-UBM framework is similar to the conventional GMM-UBM system [1]. First an ASM-UBM system prepares a universal background ASM. Then multiple sets of speaker-specific ASM are obtained by adapting the universal background ASM with enrollment data from each speaker. During testing, the front-end processing module converts the testing utterances into feature vectors. Finally the ASM sets for a claimed speaker and UBM produce a decision score for speaker recognition. A previous attempt of phonetic GMM shares a similar motivation [9], but it requires a set of phonetically labeled training data, which is not easily available in the current setting.

3.1. Background and Speaker-specific ASMs

The universal background ASM, Λ_{UBM}, can be estimated form a development set. The enrollment data from each speaker is used to estimate speaker-specific ASMs using the universal background ASM as a seed model. To perform adaptation, all enrollment utterances are decoded into sequences of ASM units with the set of universal background ASMs. Then the decoded transcriptions are used as references to establish speaker-specific ASMs by adapting the universal background ASM with any conventional HMM-based unsupervised model adaptation algorithms, such as maximum a posteriori (MAP) [10], structural-MAP [11], and stochastic matching algorithms [12]. With several iterations of decoding and adaptation, the speaker-specific ASM set can be established.

3.2. Score Calculation

The proposed ASM-UBM framework uses a normalized log-likelihood ratio for the decision score. It is calculated in a similar way to that in the GMM-UBM system [1, 2]. For a testing utterance, O, the decision score, S, is computed as:

$$S = [\, logP(O \mid \Lambda_{Hyp.}) - logP(O \mid \Lambda_{UBM}) \,] / T, \quad (2)$$

where $\Lambda_{Hyp.}$ is the ASM set for the hypothesized speaker, and T is the number of frames in the incoming utterance.

4. SPEAKER RECOGNITION EXPERIMENTS

Next we describe the experimental setup and present our ASM-UBM speaker recognition results. A baseline GMM-UBM system was established and evaluated with the same development, enrollment and evaluation sets.

4.1. Experimental Setup

In this section we describe the database and system configurations of the ASM-UBM and GMM-UBM. We also introduce the detection cost function (DCF) and detection error tradeoff (DET) curve [13] for performance evaluation.

4.1.1. Database

We evaluated the proposed ASM-UBM framework on the one-speaker detection task with the NIST 2001 corpus [14]. The development set consists of about two hours of speech. 174 target speakers, 74 male and 100 female, were used in the test. Each speaker has roughly two minutes of enrollment data used to estimate the speaker-specific ASMs. The evaluation set includes 2038 distinct test utterances. The duration of each testing utterance varies from a few seconds to one minute, with the majority of them ranging from 15 to 45 seconds. Each test utterance is evaluated against 11 hypothesized speakers (one genuine and ten imposters). As a result, the evaluation set yields a total of 22418 trials (2038 genuine and 20380 imposter).

4.1.2. Front-end Processing

For the front-end processing, we converted each input speech utterance into a sequence of 36-dimensional feature vectors, including 12 MFCC coefficients and their first and second order time derivatives. The features were then

passed through a RASTA filter [15]. An energy-based voice activity detection (VAD) algorithm was used to remove silence frames and kept only speech frames [16]. Finally, the feature vectors were processed by an utterance-level mean and variance normalization (MVN) procedure [16]. This front-end processing was used in both the GMM-UBM and ASM-UBM systems in the following experiments.

4.1.3. ASM-UBM System

For the ASM-UBM system, we used the development set to establish a set of universal background ASMs using the training procedure described in Section 2. Then, we constructed speaker specific models following the steps presented in Section 3. MAP [10] was carried out to adapt only the mean parameters of the mixture components of the universal background ASM to generate the 174 sets of speaker-specific ASMs. In this study, the number of ASMs was fixed to six and the number of active states in one ASM was set to three. Each state was characterized by 75 Gaussians. In such ASM-UBM system, the total number of Gaussians is 1350 (6×3×75) in the background ASM.

4.1.4. Baseline GMM-UBM System

To have a fair comparison, we built a set of universal background GMMs using the same amount of (1350) Gaussian mixture components for the GMM-UBM system. With the background GMM, the enrollment data were then used to estimate 174 sets of the speaker-specific GMMs. Here we also used the MAP algorithm to adapt the mean parameters of the Gaussian mixture components.

4.1.5. Zero-mean Normalization

We incorporated zero normalization (Z-norm) to normalize the testing scores to improve the overall performance [17]. The Z-norm scheme normalizes the testing score from Eq. (2) and generates the final testing score, S_{Znorm}, by:

$$S_{Znorm} = \frac{S - \mu_{Hyp.}}{\sigma_{Hyp.}}, \qquad (3)$$

where $\mu_{Hyp.}$ and $\sigma_{Hyp.}$ are mean and standard deviation statistics of the hypothesized speaker. Each speaker's mean and deviation statistics are estimated by the training data.

4.1.6. Evaluation Measure

The NIST speaker recognition evaluation uses DCF as the primary performance measure. The DCF measure is defined as a weighted sum of false rejection (FR) probability, $Pr(FR|TA)$, and false alarm (FA) probability, $Pr(FA|IM)$:

$$\mathrm{DCF} = C_{FR} Pr(FR|TA)\, Pr(TA) + C_{FA} Pr(FA|IM) Pr(IM), \qquad (4)$$

where $Pr(TA)$ and $Pr(IM)$ are the prior probabilities of the target and imposter speakers; C_{FR} and C_{FA} are relative cost factors of FR and FA, respectively. These parameters are given as: $Pr(TA)$=0.01, $Pr(IM)$=0.99, C_{FR}=10, and C_{FA}=1 [14]. It is clear that the operating point is biased towards low FA rates. We also compared the identification performance to study their speaker distinguishing capabilities. The identification error rate (IDER) is used for performance evaluation. Each IDER value (in %) is calculated by the total misidentifications out of 2038 test utterances. Another well-known measure, equal error rate (EER), will also be reported. Moreover, we demonstrated the DET curves to show the tradeoff between FR and FA.

4.2. Experimental Result

In the following, we first compared the performance attained by ASM-UBM and GMM-UBM. Then, we integrated a bigram language model to further enhance speaker characterization for the ASM-UBM system.

4.2.1. ASM-UBM versus GMM-UBM

Fig. 1 illustrates DET curves and EERs of GMM-UBM and ASM-UBM. First we find that ASM-UBM provides a 9.95% EER reduction (from 8.34% to 7.51%) over the GMM-UBM system. Moreover by comparing the two DET curves, we can clearly see that ASM-UBM achieves better performance than GMM-UBM, especially providing lower FA with a given FR. In addition to Fig. 1, we list the IDER and minDCF measures of GMM-UBM and ASM-UBM, respectively, in the first and second rows of Table I. From the table we can observe that ASM-UBM gives clear IDER and minDCF reductions of 4.92% (from 9.96% to 9.47%) and 16.40% (from 3.78% to 3.16%), respectively. The above results verify the better capability of speaker modeling of ASM-UBM over GMM-UBM. It is noted that ASM-UBM in this paper only uses six acoustic models for the background ASM. We observe a moderate increase of computation with a decoding time of about 1.5 times that of GMM-UBM when using the same amount of Gaussians.

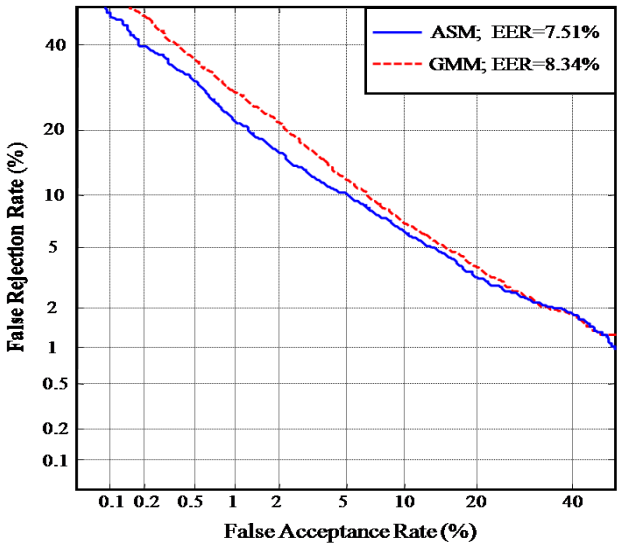

Fig. 1. DET curves of ASM-UBM versus GMM-UBM.

4.2.2. Incorporating Language Model in ASM-UBM

Next, we devised a language model (LM) for the ASM-UBM system to further take advantage of the temporal information. This LM indicates the transition properties between each pair of ASMs. With an LM, W, the target score in Eq. (2) is now calculated by: $\log P(O|\Lambda_{Hyp.}) \times P(W)$. The same criterion is applied to estimate the imposter score. In this study, we used a bigram LM that was trained on decoded ASM labels from the development set. With two hours of speech data we got a total of 47527 decoded ASM units to be used for training universal bigrams (UBM-LM). As a contrast each speaker enrollment session gave us only about two minutes of speech data or 500 decoded ASM units, which is too little to trained good speaker-specific bigrams (ASM-LM). We list this set of results as ASM-UBM+UBM-LM in the bottom row of Table I. By comparing ASM-UBM and ASM-UBM+UBM-LM, we can see that further reductions of IDER (from 9.47% to 9.27%) and minDCF (from 3.16% to 3.12%) were achieved by using the bigram LM. This confirms that LM can provide additional temporal information to characterize speakers.

TABLE I SPEAKER RECOGNITION PERFORMANCE (IN %)

System	IDER	minDCF
GMM-UBM	9.96	3.78
ASM-UBM	9.47	3.16
ASM-UBM+UBM-LM	9.27	3.12

5. CONCLUSION AND FUTURE WORK

The ASM approach resembles our recently proposed ensemble speaker and speaking environment modeling (ESSEM) framework [18]. The ASM approach attempts to establish a set of acoustic basis that spans the speech model space, while ESSEM intends to construct a set of environmental basis to span the overall speaker and speaking environment acoustic space. ESSEM was shown to enhance ASR performance under noise by characterizing the unknown environments with the ESSEM basis models. In this paper, we apply ASM to construct an ASM-UBM system for speaker recognition.

From the experimental results, we observed that the ASM-UBM system provides a clear improvement of 9.95% (from 8.34% to 7.51%) in EER over a conventional GMM-UBM system. We also found an error reduction in EER and improvements from the DET curves. The experimental results confirm that by incorporating temporal information with ASM, each speaker's acoustic characteristics can be more accurately modeled than GMM.

This paper reports our first attempt to apply ASM to speaker recognition applications. We used a fixed prototype for ASM (fixed number of ASMs and state within each ASM) and only tested MAP for speaker-specific model training. In the future we will investigate different configurations of ASM, along with other unsupervised learning algorithms, such as SMAP [11] and stochastic matching [12], for speaker-specific model training.

6. REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[2] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Proc.*, vol. 3, pp. 72–83, 1995.

[3] C.-L. Huang, B. Ma, C.-H. Wu, B. Mak, and H. Li, "Robust speaker verification using short-time frequency with long-time window and fusion of multi-resolutions," *Proc. Interspeech*, pp. 1897-1900, 2008.

[4] C.-H. Lee, F. K. Soong, and B.-H. Juang, "A segment model based approach to speech recognition," *Proc. ICASSP*, pp. 501–541, 1988.

[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected application in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[6] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 15, pp. 271-284, 2007.

[7] J. Reed and C.-H. Lee, "A study on music genre classification based on universal acoustic models," *Proc. ISMIR*, pp. 89-94, 2006.

[8] T. Svendsen and F. K. Soong, "On the automatic segmentation of speech signals," *Proc. ICASSP*, pp. 77–80, 1987.

[9] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, pp. 1969-1978, 2007.

[10] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Proc.*, vol. 2, pp. 291-99, 1994.

[11] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 276-287, 2001.

[12] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Proc.*, vol. 4, pp. 190-202, 1996.

[13] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," *Proc. Eurospeech*, pp. 1895-1898, 1997.

[14] The NIST Year 2001 Speaker Recognition Evaluation Plan. http://www.itl.nist.gov/iad/mig/tests/spk/2001/.

[15] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Proc.*, vol. 2, pp. 578-589, 1994.

[16] D. Zhu, B. Ma, and H. Li, "Joint MAP adaptation of feature transformation and Gaussian mixture model for speaker recognition," *Proc. ICASSP*, pp. 4045–4048, 2009.

[17] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing,* vol. 10, pp. 42-54, 2000.

[18] Y. Tsao and C.-H. Lee, "An ensemble speaker and speaking environment modeling approach to robust speech recognition," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 17, pp. 1025-1037, 2009.