A STUDY OF THE RELATIVE IMPORTANCE OF TEMPORAL CHARACTERISTICS IN TEXT-DEPENDENT AND TEXT-CONSTRAINED SPEAKER VERIFICATION

James H. Nealand, Jason W. Pelecanos, Ran D. Zilca, Ganesh N. Ramaswamy

Conversational Biometrics Group
IBM Research, T.J. Watson Research Center
P.O. Box 218, Yorktown Heights, NY 10598
james.n@ieee.org, {jwpeleca, zilca, ganeshr}@us.ibm.com

ABSTRACT

The relative importance of the temporal characteristics of speech for text-dependent and text-constrained speaker verification is investigated. A novel scheme is proposed using a common set of Gaussian components to form various HMM and GMM configurations, establishing a systematic transition from text-dependent to text-constrained speaker verification, and resulting in a novel alternative to conventional GMM-UBM training. Experimental results indicate that the intra-word temporal characteristics of speech do not contribute significantly to performance, however the inter-word temporal characteristics can be used during both enrollment and testing to improve verification performance.

1. INTRODUCTION

Speaker verification is the task of accepting or rejecting a claim of identity from an individual using their voice. Approaches to speaker verification may be text-dependent, text-independent or text-constrained. Text-independent methods are based on short-time analysis and ignore the textual content of the utterance. The Gaussian Mixture Model (GMM) [1] is the dominant classification architecture for text-independent speaker verification. If the expected vocabulary of a text-independent system is restricted then the system is said to be text-constrained; an example being the set of digits 0-9 and 'oh'. Text-dependent approaches to speaker verification use knowledge of the textual content of the utterance during classification. The speech signal is segmented into acoustic or linguistic classes either explicitly by a coupled speech recognition engine, by the expected speech (ie. a password or PIN), or by the verification system implicitly labelling the utterance. Hidden Markov Models (HMMs) are built for each acoustic class and aligned in some manner according to the speech segmentation [2].

This paper compares the relative importance of the temporal characteristics of cepstral features for text-dependent and text-constrained speaker verification. The effects of progressively removing the temporal selectivity of the HMM are examined. By this process, a systematic transition between HMM text-dependent and GMM text-constrained speaker recognition is established.

Experiments are conducted both where a transcription is provided, and where the classifier implicitly labels the data. The study yields an insight into the importance of temporal characteristics for speaker verification, and identifies a potential alternative to conventional GMM-UBM training.

Previously Zhu, et al. [3] compared text-independent GMM and text-dependent HMM approaches observing that the HMM method outperformed the GMM method. In another study Yu, et al. [4] observed that verification performance was primarily a function of the total number of mixture components, suggesting that the impact of intra-word temporal characteristics on verification performance is small. In these previous studies, the different HMM and GMM configurations were constructed separately making it difficult to assess the importance of the temporal selectivity of the Gaussian components.

This study is novel in that the same set of Gaussians from the originally trained HMM are used in each of the approaches. The difference between the systems is the temporal selectivity governed by the configuration of the Gaussians in the speech models.

2. TEXT-INDEPENDENT AND TEXT-CONSTRAINED GMM APPROACHES TO SPEAKER RECOGNITION

A GMM, λ, models the observed feature vectors, $O = \{\boldsymbol{o}_t; 1 \leq t \leq T\}$, as a weighted combination of a number of multivariate Gaussian densities (as shown in Equation 1) where $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ are the mean vector and covariance matrix of the m^{th} mixture component $\phi_m = \{\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}$.

$$p(\boldsymbol{o}_t \mid \lambda) = \sum_{m=1}^{M} w_m p(\boldsymbol{o}_t \mid \phi_m) \tag{1}$$

The expected frame-based log likelihood of the utterance given the model λ is:

$$E[\log p(\boldsymbol{o}_t \mid \lambda)] = \frac{1}{T} \sum_{t=1}^{T} \log p(\boldsymbol{o}_t \mid \lambda) \tag{2}$$

The expected log likelihood of a claimant speaker model $\boldsymbol{\lambda}_c$ is compared to the expected log likelihood of the Universal Background Model (UBM) λ_u built from the data from a large development set of speakers [1]. The claimant model is adapted from the UBM using Maximum *A Posteriori* (MAP) adaptation [1]. A claim of identity is accepted if the difference between the claimant score and UBM score is more than some predefined threshold.

Text-independent and text-constrained speaker verification are approached in much the same manner except in the text-constrained case the expected vocabulary of the UBM, enrollment and testing data is restricted. In text-independent and

text-constrained approaches all Gaussian components for a given speaker contribute to the scoring of all observations within the utterance; there is no temporal selectivity (or preference) for scoring particular Gaussian components over the utterance. In contrast, the state structure in text-dependent HMM speaker verification means that only some Gaussians contribute to the score for each feature vector; there is temporal selectivity in the scoring of Gaussian components over the utterance.

3. TEXT-DEPENDENT HMM APPROACHES TO SPEAKER RECOGNITION

The HMM is a finite state machine with the output emission probability density function of each state S_j represented as a weighted combination of a set of Gaussian probability densities. The transition probability matrix for a HMM with J states, $\boldsymbol{A} = \{a_{ij}; 1 \leq i, j \leq J\}$ determines the probability of moving from any given state to any other state. In speech applications, typically left-to-right HMMs are used where the transition probability matrix is constrained [2].

A HMM Φ is defined by a state transition matrix A, a set of M Gaussian mixture components $\{\phi_m\}$, mixture component weights $\{w_m\}$, and the state membership function $\{Q_m\}$ such that $\phi_m \in S_{Q_m}$.

In text-dependent HMM speaker verification systems a number of acoustic classes are defined that span the expected vocabulary of the task; typically words or phones. A speaker independent HMM is then constructed for each acoustic class to form a background model analogous to the UBM approach described for GMMs. Target speaker models are enrolled using MAP adaptation from the speaker independent model set.

To apply the correct HMM to each section of the utterance, the utterance must first be segmented into the defined acoustic classes. This segmentation may be performed using Automatic Speech Recognition (ASR), using additional knowledge of the expected textual content of the utterance, or by the speaker verification classifier implicitly transcribing the speech. The verification system can implicitly transcribe the speech by using the HMM set in conjunction with an appropriate grammar or language model.

Following the segmentation, the HMMs for each segment can be concatenated to form a single HMM spanning the utterance. During testing the utterance is scored using the Viterbi algorithm which determines the maximum likelihood state alignment X_t of the observed feature vectors such that $\boldsymbol{o}_t \in S_{X_t}$:

$$p(\boldsymbol{o}_t \mid \Phi, X_{t-1}) = a_{X_{t-1}X_t} \sum_{m=1}^{M} w_m p(\boldsymbol{o}_t \mid \phi_m) \delta_{Q_m, X_t} \quad (3)$$

Here δ is the Kronecker delta function. The temporal selectivity of the Gaussians in the HMM approach is governed by the state membership function, and the alignment of the observed feature vectors to specific states and acoustic classes. This paper explores the importance of temporal selectivity in the scoring of Gaussian components over an utterance. This is achieved by altering the level of temporal selectivity in a text-dependent HMM system, establishing a systematic transition between text-dependent HMM and text-constrained GMM approaches to speaker verification.

4. THE TRANSITION FROM TEXT-DEPENDENT TO TEXT-CONSTRAINED APPROACHES

A novel approach to examining the transition between text-dependent and text-independent speaker verification is proposed. The temporal selectivity of the HMM is reduced by grouping adjacent states. A grouping of J states beginning with state j, $\hat{S}_l = \{S_i; j \leq i < (j+J)\}$ is effectively a mapping of the state membership function such that $\hat{Q}_m = l \forall m \ni j \leq Q_m < (j+J)$. This process is shown in Figure 1.

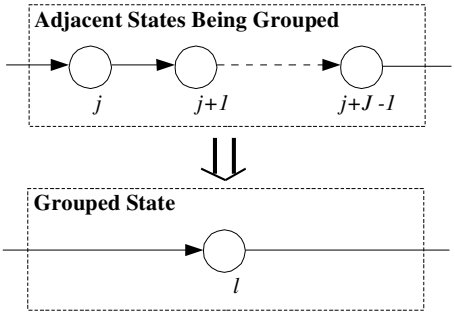

Fig. 1. *Progressive removal of temporal selectivity through grouping of adjacent states.*

The expected duration (in frames) of an utterance in a particular state is the reciprocal of the probability of exiting that state. The expected duration can be determined as Equation 4.

$$D_j = \lim_{N \to \infty} \sum_{n=1}^{N} n a_{jj}^{n-1}(1 - a_{jj}) = \frac{1}{1 - a_{jj}} \quad (4)$$

The sum of the expected duration in each member state of a group then gives the expected duration of the grouped state $\hat{D}_l = \sum_{i=j}^{j+J-1} D_i$. The expected duration of the grouped state can then be used to estimate the new transition probability matrix. The mixture component weights are also weighted such that $\hat{w}_m = \frac{D_j w_m}{\hat{D}_l}$, where $Q_m = j$ and $\hat{Q}_m = l$.

The mapping of Gaussian components from each HMM into a single GMM is equivalent to grouping adjacent states. The GMM has no transition probability matrix since there is only a single state for the entire utterance.

5. EXPERIMENT

A text-dependent word-based HMM speaker verification system was developed for telephony applications with a digit vocabulary. The speech data was collected over landline and cellular telephony channels. Three sessions of data were collected from each of the 354 enrolled target speakers. For each speaker, two instances of each digit from a single session were used for enrollment data. Approximately 2 hours of speech data were used to create the UBM. Each test utterance consisted of the claimant speaker uttering twenty digits in a random order.

The 8kHz μ-law sampled speech was first pre-emphasised with a factor of 0.97 and enframed with a window length of 20ms and frame rate of 10ms. An energy based algorithm was used to remove silence. A 24 dimension telephone bandwidth Mel-scale filter bank was applied to produce 19 cepstral features and 19 Δ

coefficients. Cepstral Mean Subtraction (CMS) was applied to reduce channel mismatch.

For the experiments described in section 5.2 the transcription of the utterance was provided for the segmentation of each utterance. For the experiments described in section 5.3, the HMM set was first used to transcribe the utterance.

5.1. Baseline Text-Dependent and Text-Constrained Approaches

A text-constrained GMM-UBM with 1056 mixture components was trained using the Expectation Maximisation (EM) algorithm. The GMM-UBM was then adapted to each target speaker using MAP adaptation. A left-to-right HMM with 6 emitting states and 16 mixture components per state was built for each digit. A UBM was first constructed using the Baum-Welch algorithm. The target models were created using MAP adaptation of the component means with a relevance factor of 15. During testing the UBM and claimant speaker models were each forcibly aligned to the test utterance using the expected transcription of the utterance. For the final GMM based system, testing of the GMM for the claimant and UBM models was conducted over the entire test utterance.

5.2. Progressive Reduction of Temporal Selectivity

Four different levels of temporal selectivity were tested. The 6 state word level HMMs were first reduced to three state models, and then reduced to single state models for each digit. The text-dependent GMMs were then combined to form a text-constrained GMM based system. These four configurations are summarised in Table 1.

Table 1. *Four approaches to speaker verification with different levels of temporal selectivity.*

System	States	Gaussians/State	Description
1	6/digit	16	Text-Dependent HMM
2	3/digit	32	Text-Dependent HMM
3	1/digit	96	Text-Dependent GMM
4	1	1056	Text-Constrained GMM

In all of the 4 approaches the total set of Gaussian components for each target speaker, and the UBM remained the same. The text-dependent GMM structure in system 3 is similar to the text constrained GMMs proposed by Sturim, et al. [5] in that the GMMs are focused on specific word classes. Figure 2 compares the performance of the 4 different levels of temporal selectivity described in Table 1, where the fusion of states is conducted after target model adaptation. The fusion of HMM states prior to target model adaptation was then examined. In this scenario the set of Gaussian densities comprising the different UBM systems remains the same, but differs between target models. Figure 3 compares the 4 different levels of temporal selectivity, with the fusion of states conducted prior to enrollment. Figures 2 and 3 suggest that there is negligible difference in performance between the text-dependent approaches with varying levels of temporal selectivity before and after enrollment. Figure 4 compares the two fused-state GMM approaches to a baseline GMM approach and shows that temporal selectivity during UBM training and target enrollment, or even UBM training alone, can benefit verification performance. Figure 4 suggests that temporal selectivity during training and enrollment each offers an incremental benefit.

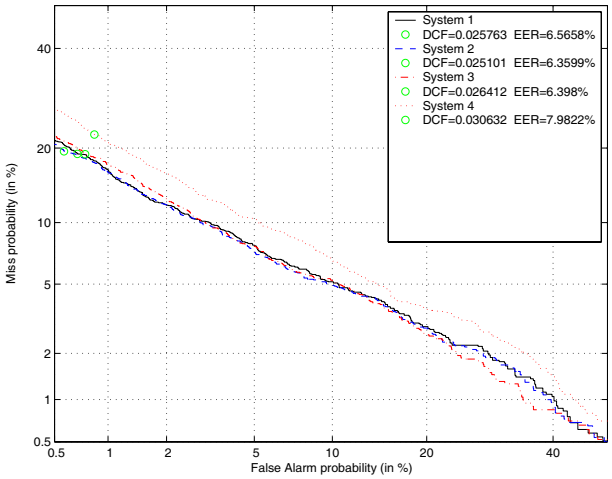

Fig. 2. *Comparison of different levels of temporal selectivity during testing.*

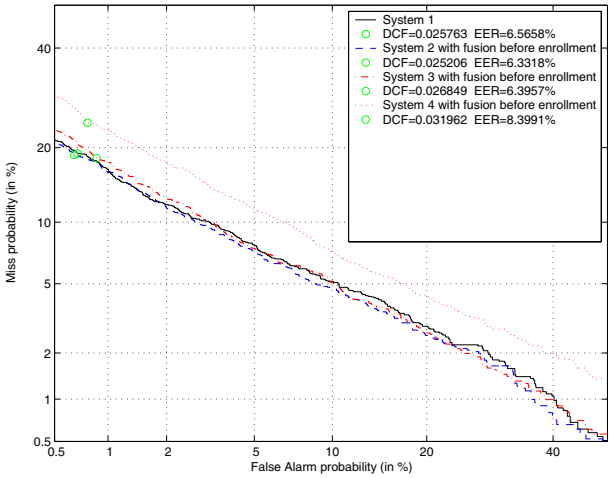

Fig. 3. *Comparison of different levels of temporal selectivity during both enrollment and testing.*

5.3. Text-Dependent HMM with Implicit Transcription

The text-dependent approaches with coupled transcriptions were then compared to a classifier that implicitly transcribes the speech. Both a 6-state per word, and a 1-state per word system were developed. The models did not utilise the actual speech transcription during either enrollment or testing. Figure 5 compares the implicit transcription approaches to the baseline HMM and GMM methods.

6. DISCUSSION

The results of Figure 2 suggest that the temporal selectivity of the HMMs, that is the number of states in the models, has negligible impact during testing when the HMMs are forcibly aligned to the labelled speech data. The text-dependent approach does however outperform the text-constrained approach suggesting the importance of inter-word temporal selectivity during testing. Figure 3

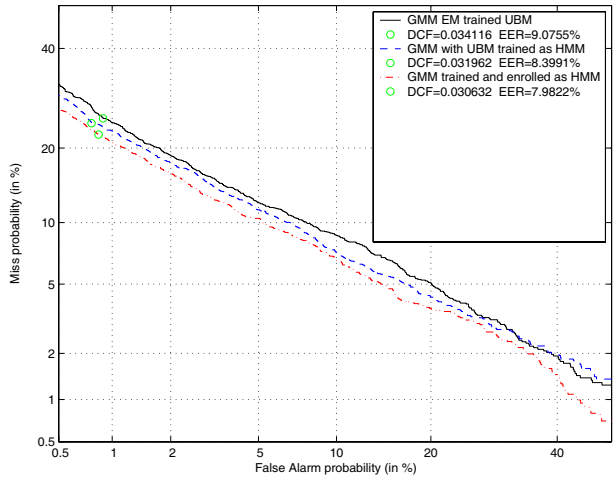

Fig. 4. *Comparison of GMM performance with temporal selectivity reduced at different stages.*

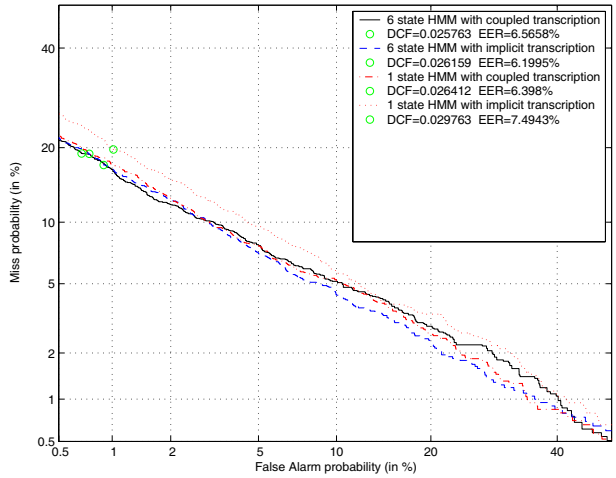

Fig. 5. *Performance of implicitly transcribed speech labels.*

is consistent with Figure 2, suggesting that for text-dependent approaches with models forcibly aligned to a coupled transcription, the intra-word temporal selectivity of the models contributes negligibly to performance during both enrollment and testing.

For the text-constrained GMM approach, Figure 4 suggests that inter-word temporal selectivity during enrollment and testing can improve performance. In the experiment, the GMM where the states are fused prior to target adaptation, is shown to outperform an equivalent conventionally trained GMM. The GMM where the states are fused after the target enrollment phase further improves performance. This indicates that the textual information in speech may be used during the UBM training, even if it is discarded during enrollment and testing. This result is of significance, since UBM training is an off-line process.

Figure 5 shows that a digits-based verification system with implicit transcription matches the performance of the forced alignment approach for 6-state word HMM systems, negating the need for a transcription during enrollment or testing. A similar observation was made for the 3-state word HMM system. This observation does not hold for the 1-state per word models where the coupled transcription outperforms the implicit transcription approach.

Figure 5 also shows that the 6-state per word system marginally outperforms the single state per word models for the implicit transcription of the data. This observation contradicts the observations from Figures 2 and 3, suggesting that the intra-word temporal information is important for the correct transcription of the speech data, however does not contribute specifically to differentiating between speakers.

7. CONCLUSIONS

Experimental evidence suggests that intra-word temporal selectivity has a negligible impact on speaker verification performance where text labels accompany the speech. The inter-word temporal selectivity does however impact performance, which accounts for the superior performance of text-dependent approaches over text-independent approaches.

A finding from this study is that by incorporating temporal selectivity of the Gaussians during the UBM training phase, improvements can be made over conventional GMM-EM training. This means that any knowledge of the textual content of the development and enrollment data can be used to benefit performance even if during testing the system will be scored as a GMM.

It is shown that a text-dependent system with an implicit scheme for transcribing the speech data is comparable to the performance of a system with pre-labelled data. Furthermore, the intra-word temporal selectivity of the word models influences verification performance when the verification system is implicitly transcribing the data.

8. REFERENCES

[1] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 91–108, 2000.

[2] ChiWei Che, Qiguang Lin, and Dong-Suk Yuk, "An HMM approach to text-prompted speaker verification," *ICASSP*, vol. 2, pp. 673–676, 1996.

[3] Xiaoyuan Zhu, Bruce Millar, and Iain Macleod et al., "A comparative study of mixture-Gaussian VQ, ergodic HMMs and left-to-right HMMs for speaker recognition," *Int. Symposium on Speech, Image Processing*, pp. 618–621, 1994.

[4] K. Yu, J. Mason, and J. Oglesby, "Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation," *IEE Proceedings on Vision, Image and Signal Processing*, vol. 142, no. 5, pp. 313–318, 1995.

[5] D.E. Sturim, D.A. Reynolds, R.B. Dunn, and T.F. Quatieri, "Speaker verification using text-constrained Gaussian mixture models," *ICASSP*, vol. 1, pp. 677–680, 2002.