INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Text Dependent Speaker Verification Using Un-supervised HMM-UBM and Temporal GMM-UBM

Achintya Kr. Sarkar and Zheng-Hua Tan

Department of Electronic Systems, Aalborg University, Denmark
akc@es.aau.dk, zt@es.aau.dk

Abstract

In this paper, we investigate the Hidden Markov Model (HMM) and the temporal Gaussian Mixture Model (GMM) systems based on the Universal Background Model (UBM) concept to capture temporal information of speech for Text Dependent (TD) Speaker Verification (SV). In TD-SV, target speakers are constrained to use only predefined fixed sentence/s during both the enrollment and the test process. The temporal information is therefore important in the sense of utterance verification, i.e. whether the test utterance has the same sequence of textual content as the utterance used during the target enrollment. However, the temporal information is not considered in the classical GMM-UBM based TD-SV system. Moreover, no transcription knowledge of the speech is required in the HMM-UBM and temporal GMM-UBM based systems. We also study the fusion of the HMM-UBM, the temporal GMM-UBM and the classical GMM-UBM systems in SV. We show that the HMM-UBM system yields better performance than the other systems in most cases. Further, fusion of the systems improve the overall speaker verification performance. The results are shown in the different tasks of the RedDots challenge 2016 database.

Index Terms: Un-supervised HMM-UBM, Temporal GMM-UBM, Text Dependent, Speaker Verification

1. Introduction

Speaker Verification (SV) is the task of either accepting or rejecting a claimant (person) by using his/her voice. It is broadly classified into two categories: Text Independent (TI) and Text Dependent (TD). In the TI system, speakers/users can speak any sentence to deliver a voice sample during the enrollment (training) and test phases of the system. In the case of the TD, speakers/users are constrained to speak a particular sentence during the test phase, which is pre-defined in the enrollment process. It is well known that text dependent system provides higher accuracy in speaker verification than a text independent; since the TD system uses the same text/pass phrases during testing which phonetically match the enrollment phrases.

In practice, real-life applications impose a constraint on the amount/duration of data which can be used for target speaker training and testing. Generally, it is expected that the test utterance will be very short (1-2s). The performance of the speaker verification system degrades significantly when training and test data are very short [1, 2].

The paper reflects some results from the OCTAVE Project (#647850), funded by the Research European Agency (REA) of the European Commission, in its framework programme Horizon 2020. The views expressed in this paper are those of the authors and do not engage any official position of the European Commission.

In recent years, text dependent speaker verification using short utterances has attracted a great interest in the research community. Many techniques have been introduced in literature to improve the text dependent SV system by capturing the phonetic temporal information from a speech signal in different modeling paradigms namely, Deep Neural Network (DNN) [3, 4], i-vector [3, 5], Hierarchical multi-Layer Acoustic Model (HiLAM) [6, 7], phone-dependent Hidden Markov Model (HMM) [8, 9] and domain adaptation [10] concepts. In [3], phonetic information is incorporated into an i-vector system by accumulating statistics from speech with respect to a pre-defined phonetic class-specific DNN output node. In [4], the intermediate output of the DNN layers are used to vectorize characterization of speech data in TD-SV. HiLAM builds a HMM model by concatenating the segmented utterance-wise models adapted from the Gaussian Mixture Model- Universal background Model (GMM-UBM) [7]. In domain adaptation [10], the mismatch between the text independent and the text dependent data is reduced by transforming the text-independent data to better match the text-dependent task (using the a-priori transcription knowledge of the text dependent data). In conventional HMM based TD-SV systems [8, 9], phoneme (context) dependent speaker models are built using the knowledge of speech transcriptions. All of these techniques depend on transcriptions of speech data either obtained by Automatic Speech Recognition (ASR) [3, 4, 8, 9, 10] or from the text phrase content of the target speaker training data [6] for TD-SV.

In this paper, we investigate an un-supervised HMM-UBM and temporal GMM-UBM based system to capture the temporal information available in the speech signal for TD-SV *without any knowledge of the speech transcriptions*. In the first approach, a multi-state Speaker Independent (SI) HMM is built *without using any transcriptions* of the speech data as label information *called HMM-UBM*, where a single dummy word (e.g. "HELLO") is forced as the transcription to all speech data during the HMM training. The state transition model parameters of HMM [11] will capture the global speaker independent temporal (phonetic) information available within the training data. This information is basically not accounted in the conventional GMM-UBM based text dependent speaker verification system. Finally, Speaker Dependent (SD) HMM models are derived from the HMM-UBM with Maximum a Posteriori (MAP) adaptation using their corresponding training data in the enrollment phase. In the test phase, the test utterance is forced aligned to the claimant HMM and HMM-UBM models for log likelihood ratio calculation. We call it the *un-supervised HMM-UBM* based SV system.

In case of a temporal GMM-UBM based method, we try to capture the target speaker specific temporal information by calculating transition probability among the GMM-UBM mixture

http://dx.doi.org/10.21437/Interspeech.2016-362

components using his/her training data. The transition probability between the particular two mixtures is calculated based on the number of adjacent frames hard aligned after a (say, i) Gaussian to other (say, j) Gaussian. In a test phase, speaker specific transition probability among the GMM-UBM components are incorporated during the log likelihood ratio calculation between the claimant and the GMM-UBM. We call it the *Temporal-GMM-UBM (TEP-GMM-UBM)* system.

Finally, we also study the fusion of the systems in the score domain. We show that the un-supervised HMM-UBM system shows better speaker verification performance than the baseline and the TEP-GMM-UBM methods for target-/imposter-wrong types in most cases. However, fusion of the systems further improve the performance of the speaker verification.

For the baseline, we consider the conventional GMM-UBM based speaker verification system. We observe that straight forward application [6] of the i-vector technique does not yield better or equivalent performance compared to the classical GMM-UBM based SV system in the RedDots challenge database (consisting very short utterance). Besides, the proposed method does not use any transcriptions of speech data, thus we restrict ourselves to a GMM-UBM based system as the baseline.

The paper is organized as follows: Section 2 & 3 describe the un-supervised HMM-UBM and TEP-GMM-UBM methods, respectively. Section 4 describes the baseline system. Experimental setup is presented in Section 5. Section 6 presents the results and discussion. Finally, the paper is concluded in Section 7.

2. Un-supervised HMM-UBM SV method

A HMM-UBM model is built using data from many non-target speakers *without any knowledge of the speech transcriptions*. So a dummy word is assigned as a (forced) transcription label (e.g. "HELLO") for *all training data* during the HMM training as shown in Fig.1. HMM-UBM is initialized with flat start and then the parameters are re-estimated with few iterations of Baum-Welch algorithm. Since we are not using any transcriptions of the training data, state transition probabilities of the HMM [11] will inherently capture/reflect the global speaker independent temporal information available within the data. This temporal information is not considered in the conventional GMM-UBM based text dependent SV system. During the enrollment phase, Speaker Dependent (SD) models are derived from the HMM-UBM with MAP adaptation [12] using his/her training data.

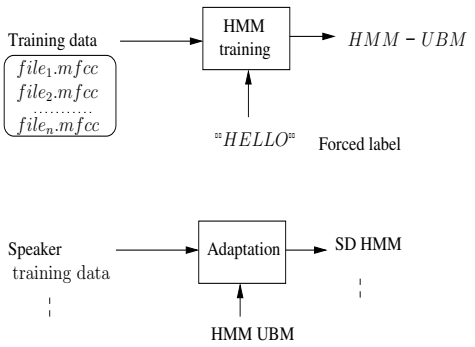

Figure 1: *Training of un-supervised speaker independent HMM-UBM and speaker dependent model without any transcription of speech data.*

In testing, the test utterance is forced aligned against the claimant-HMM and the HMM-UBM for Log Likelihood Ratio (LLR) calculation as,

$$LLR_{hmm}(X) = \frac{1}{T}\{\log p(X|q^{tar}, \lambda_{tar-hmm}) - \log p(X|q^{hmm-ubm}, \lambda_{hmm-ubm})\}$$

where $X = \{x_1, x_2, \ldots, x_T\}$ represents the feature vectors of the test utterance. $\lambda_{tar-hmm}$ and $\lambda_{hmm-ubm}$ denote the claimant and HMM-UBM, respectively. q^{tar} and $q^{hmm-ubm}$ denote the state sequence with respect to the $\lambda_{tar-hmm}$ and $\lambda_{hmm-ubm}$ models, respectively for the given test data.

3. Temporal GMM-UBM SV method

In this technique, our motivation is to capture the *target speaker specific* temporal information with respect to the Gaussian components of the GMM-UBM using his/her training data at the enrollment phase. The *speaker specific* temporal information is calculated in terms of transition probabilities among the Gaussian mixtures in the GMM-UBM. For that, the training data of the particular target speaker is first assigned to the Gaussian components of the GMM-UBM at the frame level with a hard-decision (based on maximum posteriori). Then, the transition probability between the two particular Gaussians is calculated based on the frame count. *Algorithm* 1 explains the estimation of transition probabilities with respect to the Gaussian components of the GMM-UBM $\sim \mathcal{N}(w, \mu, \Sigma)$ for the r^{th} speaker training data $X = \{x_1, x_2, \ldots, x_T\}$.

Algorithm 1: Estimate transition probability

Step 1: Estimate posteriori alignment of feature vector X with respect to the GMM-UBM

$$p(j|x_t) = \frac{w_j b_j(x_t)}{\sum_{k=1}^{M} w_k b_k(x_t)} \quad (1)$$

Step 2: Align the frames to 1-best Gaussian

$$\hat{k} = \arg \max_{1 \le j \le M} p(j|x_t) \quad (2)$$

e.g., Gaussian index of frames

$$\{10, 1, \ldots\}$$

Step 3: Count the number of frames hard-assigned after i^{th} mixture to j^{th},

$$\#frames\ (i \rightarrow j) \quad (3)$$

Step 4: Calculate transition probability between i & j mixtures,

$$a_{ij}^{r} = \frac{\#frames\ (i \rightarrow j)}{T} \quad (4)$$

Step 5: Repeat Step 3 to 4 for all combination of mixtures

Transition probabilities are calculated with respect to the fixed sequence of Gaussian mixtures in GMM-UBM for all speakers. Further, we only consider the left to right transition and hence self transition probability (within the Gaussian component) is not accounted except for the initial mixture i.e., a_{11}.

The likelihood calculation in this method can be defined using Eq.(5) for a given *single feature vector* x_t and transition probability a with respect to the GMM-UBM $\lambda_{ubm} \sim \mathcal{N}(w, \mu, \Sigma)$, and is illustrated in Fig.2.

$$\tilde{p}(x_t|\lambda_{ubm}, a) = a_{11}w_1b_1(x_t) + \sum_{i=1}^{M-1} a_{ii+1}w_{i+1}b_{i+1}(x_t) \quad (5)$$

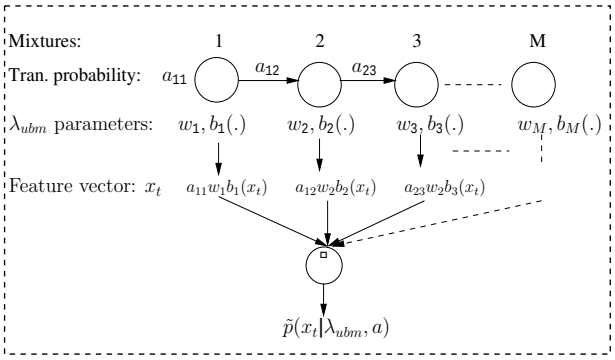

Figure 2: *Illustration of likelihood calculation for a given single feature vector x_t, GMM-UBM λ_{ubm} and transition probability a in TEM-GMM-UBM technique.*

We can rewrite the Eq.(5)

$$= \frac{a_{11}w_1}{\beta}b_1(x_t) + \sum_{i=1}^{M-1} \frac{a_{ii+1}w_{i+1}}{\beta}b_{i+1}(x_t) \quad (6)$$

$$= \tilde{w}_1b_1(x_t) + \sum_{i=1}^{M-1} \tilde{w}_{i+1}b_{i+1}(x_t) \quad (7)$$

$$= \sum_{i=1}^{M} \tilde{w}_ib_i(x_t) \quad (8)$$

where $\beta = a_{11}w_1 + \sum_{i=1}^{M-1} a_{ii+1}w_{i+1}$ is a scaling factor and Eq.(8) satisfies the GMM property $\sum_{i=1}^{M} \tilde{w}_i = 1$. M and $b_i(.)$ denote the number of mixtures in GMM-UBM and probability density function of i^{th} mixture, respectively. a_{ij} indicates the transition probability from mixture i to j.

In the enrollment phase, the target speaker-specific model (with MAP adaptation) and temporal information (i.e., transition probability) are estimated with respect to the GMM-UBM using his/her training data.

In the test phase, the log likelihood ratio for the test data $X = \{x_1, x_2, \ldots, x_T\}$ is calculated between the r^{th} claimant model λ_r and GMM-UBM $\lambda_{gmm-ubm}$ incorporating the transition probability a_r (obtained during the training phase) as,

$$LLR_{tran}(X) = \frac{1}{T}\sum_{t=1}^{T}\{\log \tilde{p}(x_t|\lambda_r, a_r) - \log \tilde{p}(x_t|\lambda_{gmm-ubm}, a_r)\}$$

It is noted that we only updated the Gaussian means of the GMM-UBM during the MAP adaptation, and so both the GMM-UBM and target models hold a point to point Gaussian link.

4. Baseline system

This system is based on the conventional GMM-UBM based speaker verification system. During training, target speaker models are derived from the GMM-UBM with MAP adaptation [7] using their training data.

In the test phase, the test utterance $X = \{x_1, x_2, \ldots, x_T\}$ is scored against the claimant λ_r and GMM-UBM $\lambda_{gmm-ubm}$ for log likelihood ratio calculation as,

$$LLR(X) = \frac{1}{T}\sum_{t=1}^{T}\{\log p(x_t|\lambda_r) - \log p(x_t|\lambda_{gmm-ubm})\}$$

Finally, LLR values are used to calculate the system performance.

5. Experimental setup

All experiments are performed using male speakers from the RedDots 2016 challenge database as per the evaluation plan [13]. RedDots evaluation is focused on various aspects of text dependent speaker systems in real-life scenarios. The evaluation tasks are primarily divided into four parts based on the different phonetic match/mismatch conditions of the target speakers training and test phase:

- **part-01**: training or testing sentences/texts are common across all speakers
- **part-02:** each speaker has their own unique sentences (text/pass phase) which are not common to all
- **part-03:** 2 free sentences (text) chosen by the individual speaker
- **part-04:** free text per speaker but the text is unique across their recording session

In brief, phonetic/lexical contents are varied among the speakers for applications of the text dependent speaker verification system in various situations. Each part is further divided into three subtasks based on the imposter type:

- **target wrong:** when a target speaker speaks a wrong sentence in the testing phase as compared to the enrollment phase
- **imposter correct:** the imposter speaks a correct sentence (same as target in the enrollment phase)
- **imposter wrong:** the imposter speaks a wrong sentence (differs from target in the enrollment phase)

Each utterance in the database is very short and on average 2-3s. For more details, see [13].

For spectral analysis, 57 dimensional MFCC (with RASTA [14] filtering) feature vectors consisting of static C_1-C_{19} cepstra, with Δ and $\Delta\Delta$ coefficients are extracted from the speech signal using 10 ms frame shift and a 20 ms Hamming window. An energy based Voice Activity Detection (VAD) is applied to removed the less energized frames. Then, the energized feature vector are normalized to zero mean and unit variance at utterance level.

The gender dependent HMM-UBM (14 states including start and emitting, 8 mixtures per state) and the GMM-UBM consisting of 96 Gaussian mixture components with diagonal covariance matrixes are trained using data (42325 utterances) from 157 male non-target speakers in the RSR2015 database [15]. For TEM-GMM system, GMM-UBM consisting of 8 mixtures is considered, as larger size of GMM-UBM yields

sparsity when estimating the transition probability based on frame count. During MAP adaptation, only the Gaussian means of the GMM-UBM and HMM-UBM are updated with 3 iterations. The value of relevance factor considered in the MAP is 10. HMM-UBM and GMM-UBM systems are implemented using HTK toolkit [16].

System performance is evaluated in terms of Equal Error Rate (EER) and Minimum Detection Cost Function (MinDCF) as per NIST 2008 SRE plan [17].

6. Results and Discussion

Tables 1- 4 compare the Text Dependent (TD) Speaker Verification (SV) performance of the different systems in various parts (tasks) of the RedDots challenge 2016 for different types of imposters.

It is observed from Tables 1-4 that the un-supervised HMM-UBM (total 96 mixtures) system shows lower error rate (either in terms of EER or MinDCF) compared to the un-supervised TEP-GMM-UBM (GMM-UBM of 8 mixtures) and baseline (GMM-UBM of 96 mixtures) systems for target-/imposter-wrong types in most cases. Results indicate that the HMM-UBM system is able to capture the speaker independent temporal information which is helpful to reject more better target-/imposter-wrong in TD-SV. Moreover, the HMM-UBM system will be also useful when needed to verify/cross-check whether target speakers are delivering the correct pass-phrase during the recording of their enrollment data.

The error rate (either in terms of EER or MinDCF) of the TEP-GMM-UBM is significantly higher than the other systems. This could be due to the fact that limited amount of training data per target makes sparse estimation of the transition probability among the GMM-UBM Gaussian components. However, fusion of the baseline with HMM-UBM and un-supervised TEP-GMM-UBM again reduces the EER and the MinDCF with respect to the standalone baseline system in most cases. It indicates that the all systems contain complementary information for the others and is useful for TD-SV.

When combining the different systems, weighted fusion is applied. The weights are estimated in several steps in the respective task: (1) an average EER value is calculated per system across the different non-target types (on evaluation set), (2) average EER values of the respective systems are then divided by the summation of average EER values of all systems, (3) intermediate weights of the systems are defined by reciprocal operation of output value at Step (2), and (4) finally, weights for the systems are calculated by re-scaling the intermediate weights obtained at Step (3), such that their summation satisfies unity.

Table 1: *Comparison of speaker verification performance of the different systems on* m-part-03 *task of RedDots challenge.*

System	Non-target type [%EER/(MinDCF× 100)]		
	target-wrong	imposter-correct	imposter-wrong
1. Baseline	4.53/2.754		**0.97**/0.342
2. TEP-GMM-UBM	20.22/6.045	no trials	9.35/2.901
3. Unsup. HMM-UBM	**4.20/1.969**	available	1.29/**0.284**
4. (1,2) (fusion)	4.20/2.788		0.80/0.275
5. (1,3) (fusion)	3.88/2.192		0.97/0.241
6. (1,2,3) (fusion)	**3.88/2.127**		**0.80/0.224**

We also believe that better fusion technique (parametric) will further improve the system performance in future with compared to our simple fusion strategy, e.g. logistic regression fusion and support vector machine fusion [18].

Table 2: *Comparison of speaker verification performance of the different systems on* m-part-04 *task of RedDots challenge.*

System	Non-target type [%EER/(MinDCF× 100)]		
	target-wrong	imposter-correct	imposter-wrong
1. Baseline	5.89/2.564	**4.19/1.890**	1.52/**0.480**
2. TEP-GMM-UBM	21.07/7.376	14.36/5.662	9.62/3.385
3. Unsup. HMM-UBM	**4.89/2.077**	5.28/2.582	**1.42**/0.500
4. (1,2) (fusion)	6.12/2.660	**3.81/1.741**	1.07/0.418
5. (1,3) (fusion)	**4.69/2.086**	4.14/2.015	1.04/0.360
6. (1,2,3) (fusion)	4.79/2.125	3.93/1.901	**0.81/0.315**

Table 3: *Comparison of speaker verification performance of the different systems on* m-part-01 *task of RedDots challenge.*

System	Non-target type [%EER/MinDCF× 100]		
	target-wrong	imposter-correct	imposter-wrong
1. Baseline	5.64/2.361	**4.19/1.882**	1.72/**0.524**
2. TEP-GMM-UBM	20.54/7.106	15.15/5.721	9.34/3.569
3. Unsup. HMM-UBM	**4.56/1.880**	4.99/2.372	**1.38**/0.541
4. (1,2) (fusion)	5.73/2.454	3.91/**1.723**	1.06/0.454
5. (1,3) (fusion)	**4.50**/1.874	4.04/1.937	1.02/0.381
6. (1,2,3) (fusion)	4.54/**1.864**	**3.86**/1.851	**0.79/0.341**

Table 4: *Comparison of speaker verification performance of the different systems on* m-part-02 *task of RedDots challenge.*

System	Non-target type [%EER/(MinDCF× 100)]		
	target-wrong	imposter-correct	imposter-wrong
1. Baseline	6.61/2.868		**1.06/0.387**
2. TEP-GMM-UBM	21.14/7.562	no trials	8.86/3.032
3. Unsup. HMM-UBM	**5.78/2.376**	available	1.50/0.463
4. (1,2) (fusion)	7.12/2.991		1.19/0.394
5. (1,3) (fusion)	**5.57/2.374**		1.00/0.306
6. (1,2,3) (fusion)	5.70/2.458		**0.93/0.302**

7. Conclusion

In this paper, we investigated two un-supervised methods to capture temporal information from speech in an un-supervised manner for text dependent speaker verification. One is based on HMM-UBM and an other is TEP-GMM-UBM. In the HMM-UBM system, a speaker independent multi-state HMM is trained using data from many non-target speakers *without any knowledge of speech transcriptions* to capture the global speaker independent temporal information available within the speech. The target speaker models are then derived from the HMM-UBM using their particular training data with MAP adaptation. In the test phase, the test utterance is forced aligned to the claimant HMM and HMM-UBM for calculating the log likelihood ratio. For the TEP-GMM-UBM based technique, the target speaker specific temporal information is captured by estimating transition probability with respect to the Gaussian components of the GMM-UBM using their training data. In the test phase, the speaker specific transition probability is incorporated during the log likelihood calculation between the claimant model and GMM-UBM. We showed that the HMM-UBM system yields better performance than the baseline and TEP-GMM-UBM systems for target-/imposter-wrong types in most cases. This indicates that the speaker independent temporal information is useful for the text-dependent speaker verification. However, fusion of the systems further improve the performance of the speaker verification with respect to their standalone system. All results are presented in various tasks of the RedDots challenge 2016 database.

8. References

[1] A. K. Sarkar, D. Matrouf, P. M. Bousquet, and J. F. Bonastre, "Study of the Effect of i-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification," in *Proc. of Interspeech*, 2012, pp. 2662–2665.

[2] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector based Speaker Recognition on Short Utterances," in *Proc. of Interspeech*, 2011, pp. 2341–2344.

[3] N. Scheffer and Y. Lei, "Content Matching for Short Duration Speaker Recognition," in *Proc. of Interspeech*, 2014, pp. 1317–1321.

[4] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, "Deep Feature for Text-dependent Speaker Verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.

[5] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, pp. 788–798, 2011.

[6] A. Larcher, K. A. Lee, B. Ma, and H. Li, "RSR2015: Database for Text-dependent Speaker Verification using Multiple Pass-phrases," in *Proc. of Interspeech*, 2012, pp. 1580–1583.

[7] 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.

[8] S. Kajarekar and H. Hermansky, "Speaker Verification Based on Broad Phonetic Categories," in *Proc. of Odyssey Speaker and Language Recognition Workshop*, 2001, pp. 201–206.

[9] R. Auckenthaler, E. Parris, and M. Carey, "Improving a GMM Speaker Verification System by Phonetic Weighting," in *Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP)*, 1999, pp. 313–316.

[10] H. Aronowitz and A. Rendel, "Domain Adaptation for Text Dependent Speaker Verification," in *Proc. of Interspeech*, 2014, pp. 1337–1341.

[11] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. of the IEEE*, vol. 77, pp. 257–285, 1989.

[12] J.-L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[13] "The reddots challenge: Towards characterizing speakers from short utterances," https://sites.google.com/site/thereddotsproject/reddots-challenge.

[14] H. Hermanksy and N. Morgan, "Rasta Processing of Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 578–589, 1994.

[15] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent Speaker Verification: Classifiers, Databases and RSR2015," *Speech Communucation*, vol. 60, pp. 56–77, 2014.

[16] S. Young, D. Kershaw, J. Odell, V. Valtchev, P. Woodland, and et al., "HTK Book," *Copyright 2001-2006 Cambridge University Engineering Department*.

[17] A. Martin, G. Doddington, T. Kamm, M. Ordowskiand, and M. Przybocki, "The Det Curve in Assessment of Detection Task Performance," in *Proc. of Eur. Conf. Speech Commun. and Tech. (Eurospeech)*, 1997, pp. 1895–1898.

[18] O. Plchot et al., "Developing a Speaker Identification System for the DARPA RATS Project," in *Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP)*, 2013, pp. 6768–6772.