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

Parallel Speaker and Content Modelling for Text-dependent Speaker Verification

Jianbo Ma[1], *Saad Irtza*[1,2], *Kaavya Sriskandaraja*[1,2], *Vidhyasaharan Sethu*[1], *and Eliathamby Ambikairajah*[1,2]

[1] School of Electrical Engineering and Telecommunications, UNSW Australia
[2] ATP Research Laboratory, National ICT Australia (NICTA), Australia

jianbo.ma@student.unsw.edu.au

Abstract

Text-dependent short duration speaker verification involves two challenges. The primary challenge of interest is the verification of the speaker's identity, and often a secondary challenge of interest is the verification of the lexical content of the pass-phrase. In this paper, we propose the use of two systems to handle these two tasks in parallel with one sub-system modelling speaker identity based on the assumption that lexical content is known and the other sub-system modelling lexical content in a speaker dependent manner. The text-dependent speaker verification sub-system is based on hidden Markov models and the lexical content verification system is based on models of speech segments that use a distinct Gaussian mixture model for each segment. Furthermore, a mixture selection method based on KL divergence was applied to refine the lexical content sub-system by making the models more discriminative. Experiments on part 1 of the RedDots database showed that the proposed combination of two sub-systems outperformed the baseline system by 39.8%, 51.1% and 37.3% in terms of the 'imposter_correct', 'target_wrong' and 'imposter_wrong' metrics respectively.

Index Terms: KL divergence, hidden Markov models, segment model, RedDots database, text-dependent speaker verification, short duration speaker verification, Gaussian mixture models

1. Introduction

Automatic Speaker Verification (ASV) is a non-invasive biometric authentication technique which is used to verify the identity of speakers. There are two types of ASV systems: Text-independent speaker verification (TI) and Text-dependent speaker verification (TD). In text-independent speaker verification, speakers are free to speak anything and variability can be expected to arise due to differences in the speech content. Text Dependent systems have lexical constraints which require the speaker to speak specific pass-phrases, which are fixed prior to authentication or prompted during authentication process. Compared to TI systems, TD systems showed higher verification accuracy and short enrolment and test sessions can be employed, so that TD systems are generally preferred for security authentication scenarios [1].

Recent study of TD speaker verification focuses on the efficient modelling of speaker and lexical content information of extremely short utterances (around 1.5 seconds). However, a challenge faced by text-dependent speaker verification systems is that framing the alternative hypothesis is not straightforward. In text-independent speaker verification system, the hypotheses are straightforward. The hypothesis under test, H_χ denotes that test sentence is from the claimed speaker while the alternative hypothesis $H_{\bar{\chi}}$ is that the test sentence does not come from the claimed speaker [2]. Here, χ represents the target speaker. However, in text-dependent speaker verification system, the hypotheses under test, $H_{(\chi,\wp)}$, is that the test utterance is from the claimed speaker and the content of the utterance matches the expected pass-phrase. Consequently, there are 3 potential alternative hypotheses, namely, the speaker is not claimed speaker but the pass-phrase is right ($H_{(\bar{\chi},\wp)}$), the speaker is the claimed speaker but the test utterance is not the expected pass-phrase ($H_{(\chi,\bar{\wp})}$), and the speaker is not the claimed speaker and the test utterance is not the expected pass-phrase ($H_{(\bar{\chi},\bar{\wp})}$) [3]. These three alternate hypotheses may be referred to as imposter-correct, target-wrong and imposter-wrong respectively.

The advantages of text-dependent speaker verification over the text-independent ones arise from having prior knowledge of the pass-phrase that is to be spoken which in turn allows for the use of more accurate content specific speaker models. Recent approaches to TD speaker verification have included the generalisation of the Joint Factor Analysis (JFA) framework to consider supervector-sized $\mathcal{Z}$-vectors that model speaker-phrase combinations with promising results [6]. More recently, $\mathcal{Y}$-vectors and $\mathcal{Z}$-vectors, which are expected to characterize both speaker and pass-phrase information, were jointly used to model the left-to-right structure of utterances and a joint density backend was proposed [9]. In terms of left-to-right structure, a hidden Markov model (HMM) based system was also applied in [3] [8] [12]. In particular a hierarchical system including GMM and HMM was proposed in [3], and results showed the benefits of explicitly modelling the alternative hypothesis. In addition, deep neural network (DNN) based methods have also been used. In [10], long short-term memory (LSTM) neural network was applied as it has the ability to model temporal structure of short utterances. DNN based features were integrated into GMM-UBM framework in [11]. The GMM-UBM system showed good results which are difficult to beat in most of the TD speaker verification research [9, 11-13]. All of these approaches model both the speaker identity and the lexical content of the pass-phrase.

In this paper, we propose splitting the tasks of verifying the speaker identity and the lexical content, running two systems in parallel to handle these two tasks in parallel before combining the results. Furthermore, we introduced a mixture selection method based on KL divergence to select

http://dx.doi.org/10.21437/Interspeech.2016-825

discriminative mixtures in GMM for use in each speaker-pass-phrase model.

2. Proposed system

The proposed system comprises of two sub-systems running in parallel, one that models speaker characteristics and verifies the speaker identity operating on the assumption that the right pass-phrase was spoken and a second one that models lexical content detects if the right pass-phrase was spoken, as shown in Figure 1. Both sub-systems make use of the same front-end and the outputs of both sub-systems are combined to test against all three alternate hypotheses.

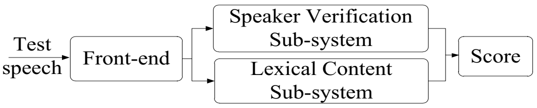

Figure 1: *Proposed parallel speaker and content modelling*

The front-end of this system comprises of Standard MFCC features of 19 dimensions with log-energy and their first and second derivatives. A vector quantization model based voice activity detector was used [16] and Feature warping [17] was applied to these features.

2.1. Speaker verification sub-system

This sub-system (denoted as λ_{nHMM}, where n is the number of states in HMM) operates on the assumption that the lexical content is known for each trial in order to verify the claimed speaker identity. It employs HMM based speaker models, where each state is represented by a suitable GMM, as shown in Figure 2. Initially, an N-state HMM is initialised with a universal background Gaussian mixture model (λ_{UBM}) in each state and retrained with all data corresponding to each pass-phrase to estimate background pass-phrase HMMs (λ_{BHMM}). Speaker specific pass-phrase HMMs (λ_{SPHMM}) are obtained via MAP adaptation off these background pass-phrase HMMs using examples of the target pass-phrase spoken by the target speaker.

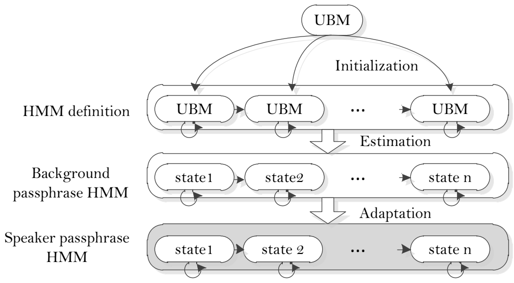

Figure 2: *Speaker verification sub-system using HMM*

For score calculation, averaged log-likelihoods $\log P(\mathcal{O}|\lambda_{\mathrm{SPHMM}})$ and $\log P(\mathcal{O}|\lambda_{BPHMM})$ for each frame in test utterance, $\mathcal{O}$, are calculated from λ_{SPHMM} and λ_{BPHMM} respectively by using Viterbi algorithm. The final score for this sub-system is formulated as:

$$\mathrm{S}_{\mathrm{HMM}} = \log P(\mathcal{O}|\lambda_{\mathrm{SPHMM}}) - \log P(\mathcal{O}|\lambda_{\mathrm{BPHMM}}) \qquad (1)$$

2.2. Lexical content sub-system

The aim of the lexical content sub-system (denoted as λ_{nseg}, where n is the number of segments in the segment modelling) is to verify that the lexical content of the test utterance matches that of the expected pass-phrase. One way to do that is to use HMM based methods as in section 2.1. However, as the number of sessions for each speaker passphrase is rather limited, using HMM based methods to estimate the state-based alignment would not guarantee accuracy. As a compromise, in the proposed system, an alternative approach utilising a left-to-right segment model is adopted.

The left-to-right segment model operates by splitting each pass-phrase into S segments and using a separate GMM to model each segment. Each segment GMM is expected to model the phonetic structure of the short segments of speech and the sequence of segment GMMs as a whole can be expected to model the overall temporal structure of the pass-phrase. i.e., two utterance that have the same content but with different phonetic order, will not generate similar scores as the order of phonemes is different.

Figure 2 shows how the left-to-right segment model is created from a suitable universal background Gaussian mixture model (λ_{UBM}). Each utterance sequence from the same pass-phrase of a particular speaker is split into S segments of equal lengths. Feature vectors from each segment used to adapt the background model, λ_{UBM}, to model that segment's lexical content and speaker information. The set of S adapted GMMs form the segment model of each pass-phrase for each speaker.

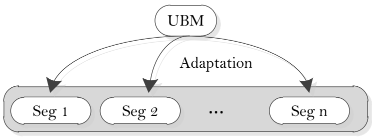

Figure 3: *Lexical content sub-system using segment models*

In the scoring phase, the test speech is divided into S segments. Each segment is scored against the corresponding segment model. The UBM is used to compute log-likelihoods from each segment. $\log P(\mathcal{O}_i|\lambda_{\mathrm{Seg(i)}})$ and $\log P(\mathcal{O}_i|\lambda_{\mathrm{UBM}})$ for each frame in test utterances are calculated. The final score is then the mean of log-likelihood ratio of each segment model:

$$S_{seg} = \frac{1}{S}\sum_{i=1}^{S}\left(\log P\left(\mathcal{O}_i\middle|\lambda_{Seg(i)}\right) - \log P(\mathcal{O}_i|\lambda_{UBM})\right) \qquad (2)$$

where, S is the number of segments, $\mathcal{O}_i$ denotes the i^{th} segment of speech and $\lambda_{Seg(i)}$ denotes the i^{th} segment GMM.

2.3. Score interpretation and combination

As previously mentioned, in text dependent speaker verification the alternative hypothesis consists of three sub-hypothesis. In the proposed system, the speaker verification sub-system estimates the log-likelihood ratio of a model of correct pass-phrase from the target speaker (λ_{SPHMM}) to a model of the correct pass-phrase from non-target speaker (λ_{BPHMM}). Consequently the sub-system score, S_{HMM} can be interpreted as comparing the hypothesis $H_{(\chi,\wp)}$ and $H_{(\bar{\chi},\wp)}$. i.e.,

$$\mathrm{S}_{\mathrm{HMM}} = \log P\left(\mathcal{O}\middle|H_{(\chi,\wp)}\right) - \log P\left(\mathcal{O}\middle|H_{(\bar{\chi},\wp)}\right) \qquad (3)$$

The left-to-right segment models used in the lexical content sub-system differ for different pass-phrases. Therefore, pass-phrases that do not share the same lexical content will lead to low likelihood values, even if they are from the same speaker, because the temporal structure is different. As a result, we assume λ_{seg} models both $H_{(\chi,\wp)}$ and $H_{(\chi,\bar{\wp})}$ and denote it as $H_{(\lambda_{\text{seg}})}$. Finally, λ_{UBM} is assumed to be text and speaker independent and the likelihood of the background model can be thought of as representing $P\left(\mathcal{O}\middle|H_{(\bar{\chi},\bar{\wp})}\right)$.

Adding scores from two sub-systems and using the interpretation above, we obtain:

$$\begin{aligned} Score = \log P(\mathcal{O}|H_{(\chi,\wp)}) - \log P\left(\mathcal{O}\middle|H_{(\bar{\chi},\wp)}\right) \\ + \log P(\mathcal{O}|H_{(\lambda_{seg})\_} - \log P(\mathcal{O}|H_{(\bar{\chi},\bar{\wp})}) \end{aligned} \tag{4}$$

Noting that $H_{(\lambda_{\text{seg}})}$ models $H_{(\chi,\wp)}$ and $H_{(\chi,\bar{\wp})}$, the combined score consists of all the three sub-hypotheses.

3. Mixture selection

In a speaker verification system, including the proposed system, the UBM is assumed to be a text and speaker independent model that covers all of the imposters and lexical content. Since each pass-phrase is extremely short and phoneme coverage is limited in short duration text-dependent speaker verification [12], it is reasonable to argue that the number of adapted mixtures in each model is quite limited, which makes the models quite redundant. Moreover, some adapted mixtures based only on a small number of feature frames can lead to errors and removing them could help. In this section, we propose the use of a Gaussian mixture selection method to select the most discriminative mixtures between UBM and adapted speaker GMM model.

A symmetric version of KL divergence (Jensen–Shannon divergence) [15] is used as a similarity measure between two Gaussian mixture models of UBM and adapted speaker model, $f(x)$ and $g(x)$ respectively. The UBM and adapted Gaussian model consists of 'M' mixtures.

$$D(f,g) = \tfrac{1}{2}[D(f||g) + D(g||f)] \tag{4}$$

Here $D(f||g)$ and $D(g||f)$ is the KL divergence between probability density function *'f'* to *'g'* and *'g'* to *'f'* respectively. The KL divergence between two mixtures (Gaussian mixtures are assumed to have diagonal covariance matrix) is defined as in [15].

$$D(f||g) = \tfrac{1}{2}\left[(w_f - w_g)\log\left(\frac{w_f}{w_g}\right) + \textstyle\sum_{i=1}^{n}\tfrac{1}{2}(w_f\sigma_{fi}^2 - w_g\sigma_{gi}^2)\left(\frac{1}{\sigma_{gi}^2} - \frac{1}{\sigma_{fi}^2}\right) + \textstyle\sum_{i=1}^{n}\tfrac{1}{2}\left(u_{gi} - u_{fi}\right)^2\left(\frac{w_f}{\sigma_{gi}^2} + \frac{w_g}{\sigma_{fi}^2}\right) + (w_f - w_g)(\tfrac{1}{2}\textstyle\sum_{i=1}^{n}(log\sigma_{gi}^2 - log\sigma_{fi}^2)\right. \tag{5}$$

Where n is the feature dimension; $w_f$and $w_g$are the weights; σ_{fi} and σ_{gi} are diagonal elements of covariance; u_{fi} and u_{gi} are elements of means of two mixtures. The $M * M$ KL divergence matrix is computed using equation-5.

As both UBM and speaker GMM have M mixtures, KL divergences between each mixture in UBM and speaker GMM will be calculated, which results in a $M * M$ KL divergence matrix. This is because KL divergence is used to measure the distance between two distributions. The larger the KL, the further the two distributions are. Discriminative mixtures are chosen based on this idea. First, the minimum element in the matrix is selected and the column index which indicate the mixture place in UBM and raw index which indicate the mixture place in GMM of this element in matrix are recorded in two vectors V_{ubm} and V_{gmm}. Next, all of the elements in this column and raw are removed as these two mixtures have been selected. This process is repeated until all of the elements are selected and two vectors have recorded all of the mixtures in ascending order. The higher the mixture in V_{ubm}or V_{gmm}, the more discriminative it is and will be selected as a discriminative mixture. Mixtures in the bottom list of V_{ubm} and V_{gmm}will be selected according to the required number of mixtures.

In the rest of the paper, systems with mixture selection will be denoted as λ_{MSm}, where m is the required number of mixtures. As mixture selection will be applied to the lexical content sub-system, λ_{MSm_nseg} will denote this system.

4. Baseline system

The baseline system is a GMM-UBM system. Standard MFCC features of 19 dimensions with log-energy and their first and second derivatives were used. A vector quantization model based voice activity detector was used [16]. Feature warping [17] was applied. A gender-dependent universal background models (UBMs) of 512 Gaussian mixtures was created using all utterances from male speakers from RSR2015 database[8]. The MFCC feature extraction and UBM training were done using the HTK toolkit [18]. The UBM is then adapted to each pass-phrase of speakers using *maximum a posterior* (MAP) algorithm by corresponding enrolment utterances. Only means of GMM are adapted. Weights and covariances are shared across all models. In the rest of the paper, the baseline system is referred as λ_{GMM}.

5. Experimental results

Experiments were conducted on the RedDots database [19]. This database is collected for short duration text-dependent speaker verification. It consists of four parts. Part 1 is Common Pass-Phrase in which every speaker has the same ten pass-phrases; In Part 2, it is unique pass-phrases text-dependent. Every speaker has 10 different sentences and there is no common sentence between speakers. In part 3, each speaker has two free-choice sentences. Part 4 contains free text sentences that are unique across all sessions. Only Part 1 was considered in this work and only results on the male condition are reported in this paper. Test protocol were provided along with the RedDots database [19]. Results are reported for three different kinds of non-target trials (imposter_correct, target_wrong, and imposter_wrong) in terms of Equal Error Rate (EER).

5.1. Parallel speaker and content modelling systems

A number of experiments using the two sub-systems described in section 2 were carried out and the results are summarised in Table 1. When using 4-segment left-to-right segment models, target_wrong is improved substantially (45.6% relative improvement). This supports the assumption that the sub-hypothesis $H_{(\chi,\bar{\wp})}$ is modelled by segment modelling in section 2.2. However, the results of imposter_correct are degraded slightly and imposter_wrong is almost the same. This is not unexpected since the segment models have no mechanism of modelling the sub-hypothesis $H_{(\bar{\chi},\wp)}$. A model that takes this sub-hypothesis into consideration should be proposed. We also used 8-segment left-to-right segmental models, but the results are not better than those obtained with the 4-segment models. This may be because the extremely short duration utterances contain limited phonemes, and therefore, having a large number of segments becomes less useful.

Table 1 also reports the results obtained with the HMM based speaker verification sub-system which is described in section 2.1. Experiments with 4 and 8 states (while keeping the total number mixtures in the HMM a constant) were carried out. The results showed that by using more states, the performance is slightly improved. Compared with the baseline λ_{GMM}, the result of imposter_correct is improved by 50%, while the results of target_wrong and imposter_wrong are degraded. This is due to the HMM based system being designed to model the sub-hypothesis $H_{(\bar{\chi},\wp)}$ only. When different pass-phrases occur in enrolment and test, both the background HMM and speaker pass-phrase HMM are mismatched in terms of content information, which means the other two sub-hypothesis are not taken into consideration by this sub-system.

As we can see from above two individual experiments, HMM based and segments models are complementary in terms of modelling the complete alternative hypothesis which contains three sub-hypothesis. Thus, it is natural to combine these two sub-systems. As the combined system models complementary alternative hypothesis, we expect it to perform better than the baseline across all three metrics. The column with the notation $\lambda_{\text{8HMM}} + \lambda_{\text{4seg}}$ lists the results when combining the two sub-systems. As analysed in section 2.3, we combined the scores from different systems and the complete alternative hypothesis is the summation of the log-likelihood of the three competing sub-hypothesis. We can see from the results, compared with the baseline, 26.7%, 46.2% and 22% relative improvement were obtained.

Table 1. *Performance (EER (%)) of speaker verification sub-system and lexical content sub-system with different states and segments on part 1 of RedDots (male part).*

$\overline{(\chi,\wp)}$	λ_{GMM}	λ_{4seg}	λ_{8seg}	λ_{4HMM}	λ_{8HMM}	λ_{8HMM} $+ \lambda_{4seg}$
$(\bar{\chi},\wp)$	2.41	2.81	5.64	1.20	1.19	**1.76**
$(\chi,\bar{\wp})$	5.11	2.78	6.29	6.42	5.92	**2.72**
$(\bar{\chi},\bar{\wp})$	0.59	0.62	2.22	1.23	1.20	**0.46**

5.2. Mixture selection

Mixture selection was conducted by using the method introduced in section 3.1. Table 2 shows the results of baseline and mixture selection. We can see that when only half of the mixtures were chosen, performances of three different types of imposters are improved. We also observe that non-target trials with wrong content will be better identified by the selected mixtures, even if the number of mixtures decreases down to 64, and the results for imposter_correct start to degrade below 128 mixtures. This observation suggests that information about lexical content can be represented by a limited number of discriminative mixtures (e.g. 64 compared with 512). This means that even though there are only a few frames aligned to a component, it may be discriminative in terms of speaker verification. This is likely to happen if we think that there is limited speaker information in short duration utterances. When the number of mixtures falls to 32, performances are degraded for all three kinds of imposters.

From the results of lexical content sub-system, it can be seen that 4 segments is better. We applied mixture selection on this system to use half of number of mixtures in each model. Table 2 in column with the notation $\lambda_{\text{4seg_ms}}$shows the results. Compared with results without mixture selection, improvements across three metrics are obtained. Further combination with the speaker verification sub-system obtained 39.8%, 51.1% and 37.3% relative improvements, which are the best results we have across the experiments.

Table 2. *Performance (EER (%)) of mixture selection with various mixtures on part 1 of RedDots (male part).*

$\overline{(\chi,\wp)}$	λ_{GMM}	λ_{MS256}	λ_{MS128}	λ_{MS64}	λ_{MS32}	λ_{MS256_4seg}	$\lambda_{\text{MS256\_}4seg}$ $+ \lambda_{8HMM}$
$(\bar{\chi},\wp)$	2.41	2.34	2.50	2.96	4.34	2.80	**1.45**
$(\chi,\bar{\wp})$	5.11	4.50	3.98	4.18	5.62	2.50	**2.50**
$(\bar{\chi},\bar{\wp})$	0.59	0.48	0.52	0.77	1.24	0.56	**0.37**

6. Conclusions

In this paper, we have proposed the use of two separate sub-systems, based on hidden Markov models and sets of segment GMMs, to model the combined speaker and lexical content information in parallel for short duration utterances. The novel lexical content sub-system detects if the right pass-phrase was spoken. The use of a mixture selection method on this sub-system was shown to be beneficial when selectively using discriminative mixtures. The performances of the individual sub-systems and that of the combined system have been evaluated on the RedDots database and the two sub-systems are shown to be complementary.

7. References

[1] M. Hébert, "Text-dependent speaker recognition," in *Springer handbook of speech processing*, ed: Springer, 2008, pp. 743-762.

[2] 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.

[3] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Modelling the alternative hypothesis for text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 734-738.

[4] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13,* 2005.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 788-798, 2011.

[6] P. Kenny, T. Stafylakis, P. Ouellet, and M. J. Alam, "JFA-based front ends for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 1705-1709.

[7] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7649-7653.

[8] A. Larcher, K.-A. Lee, B. Ma, and H. Li, "RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases," in *INTERSPEECH*, 2012, pp. 1580-1583.

[9] P. Kenny, T. Stafylakis, J. Alam, and M. Kockmann, "JFA modeling with left-to-right structure and a new backend for text-dependent speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 4689-4693.

[10] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-End Text-Dependent Speaker Verification," *arXiv preprint arXiv:1509.08062,* 2015.

[11] 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.

[12] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication,* vol. 60, pp. 56-77, 2014.

[13] P. Kenny, T. Stafylakis, J. Alam, P. Ouellet, and M. Kockmann, "Joint factor analysis for text-dependent speaker verification," in *Proc. Odyssey Speaker and Language Recognition Workshop, Joensuu, Finland*, 2014.

[14] T. Hasan and J. H. Hansen, "A study on universal background model training in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 1890-1899, 2011.

[15] Y. Lei and J. H. Hansen, "Dialect classification via text-independent training and testing for arabic, spanish, and chinese," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 85-96, 2011.

[16] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *ICASSP*, 2013, pp. 7229-7233.

[17] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," 2001.

[18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu*, et al.*, *The HTK book* vol. 2: Entropic Cambridge Research Laboratory Cambridge, 1997.

[19] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. van Leeuwen*, et al.*, "The RedDots Data Collection for Speaker Recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.