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

Joint Speaker and Lexical Modeling for Short-Term Characterization of Speaker

Guangsen Wang, Kong Aik Lee, Trung Hieu Nguyen, Hanwu Sun, Bin Ma

Institute for Infocomm Research, A*STAR, Singapore
{wang-g,kalee,mabin,hwsun,thnguyen}@i2r.a-star.edu.sg

Abstract

For speech utterances of very short duration, speaker characterization has shown strong dependency on the lexical content. In this context, speaker verification is always performed by analyzing and matching speaker pronunciation of individual words, syllables, or phones. In this paper, we advocate the use of hidden Markov model (HMM) for joint modeling of speaker characteristic and lexical content. We then develop a scoring model that scores only the speaker part rather than the joint speaker-lexical component leading to a better speaker verification performance. Experiments were conducted on the text-prompted task of RSR2015 and the RedDots datasets. In the RSR2015, the prompted texts are limited to random sequences of digits. The RedDots dataset dictates an unconstrained scenario where the prompted texts are free-text sentences. Both RSR2015 and RedDots datasets are publicly available.

Index Terms: Text-Prompted, Text-Dependent, Speaker Verification, Speaker Adaptation, Hidden Markov Model

1. Introduction

Speech utterances are acoustic realizations of word sequences [1]. The lexical content manifests itself and conflates with the vocal characteristic of the person uttering it. For speaker recognition, we are interested in the speaker characteristic while taking all other factors including the lexical content and channel effects as nuisance attributes [2]. For ease of modeling, it is always assumed that the speaker characteristic is independent of the spoken content [3, 4]. Such assumption generally holds when sufficiently long utterances are given. In this context, models like joint factor analysis (JFA) [3] and the total variability model [5], have shown to be extremely effective.

For speech utterances of very short duration (i.e., in the order of few seconds), speaker characteristic has shown significant dependency on the lexical content [6, 7, 8]. It is therefore beneficial to model both speaker and spoken content jointly, for instance, modeling of speaker pronunciation of individual words, syllables, or phones. This marks the major difference between text-dependent (TD) [6] and text-independent (TI) [2] models for speaker verification. The main idea of the former is to directly exploit the voice individuality associated with a specific lexical context. In this paper, we advocate similar form of joint speaker-lexical modeling approach for text-prompted speaker verification.

Recently, phoneme adaption scheme [9] has been used under the JFA framework for the digit-prompted speaker verification task, where a phonetic independent universal background model (UBM) was adapted by maximum *a posteriori* (MAP) [10] adaptation to characterize speaker and digits. Instead of relying on a single UBM, we choose to use a phonetic-dependent model, where hidden Markov model (HMM) is used to cover the lexical variability. Speaker adaptation is then performed by adapting the phonetic model to a speaker-dependent (SD) phonetic model. During test, a composite HMM is constructed to score the test utterance under the lexical context given by the prompted text. Notice that our intention here is different from the automatic speech recognition (ASR) task where the objective is to produce word sequence as the output. To this end, we look into the speaker verification score where our aim is to evaluate only likelihood of the speaker component rather than the speaker-lexical components due to the joint modeling. In particular, we show that the likelihood score associated with lexical content could be suppressed by discarding the state transition probabilities followed by normalization with respect to a speaker-independent (SI) phonetic model. Moreover, we also look into strategy to reject target trials with wrong text using just the likelihood score.

Compared to TD task, text-prompted task is more challenging in which user is prompted to provide utterance of random text every time the system is used. The prompted texts could be random sequences of keywords from a constrained set (e.g., digits) or unconstrained free-text sentences. In this paper, we aim to investigate both scenarios using two publicly available datasets. For the former, experiment was conducted on Part III of RSR2015 database [6]. For the unconstrained scenario, experiment was conducted on the newly acquired RedDots dataset [11]. Notably the RedDots dataset is the formal dataset for the INTERSPEECH 2016 RedDots challenge.

The remaining of the paper is organized as follows. Section 2 details the joint speaker and lexical modeling. Various scoring schemes are discussed in Section 3. Experimental results on the RSR2015 and RedDots corpora are presented in Section 4 and Section 5 respectively. Section 6 summarizes the findings and concludes the paper.

2. Speaker and lexical modeling

2.1. Speaker independent model

Lexical content can be represented at a number of levels, from words to syllables, phones or below [1]. Among these, phone, or more precisely tri-phone is the most common choice used in modern speech recognition system. These acoustic modeling units are usually modeled using HMMs. Due to the physiological limitations in the speech production of individual speaker, the realizations of the same acoustic unit are never the same. In this regard, the transition between HMM states accounts for the temporal variance, while the state probability distribution models the acoustic variation.

Let N be the number of phones, for instance, $N = 40$ is typical for English. Considering the left and right contexts, the number of tri-phone HMM models is N^3. With a proper ty-

http://dx.doi.org/10.21437/Interspeech.2016-929

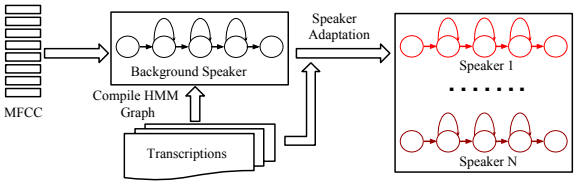

Figure 1: *Speaker enrollment via MAP*

ing of states across the HMMs, the number of tied sates (i.e., senones) can be reduced considerably to be a few hundreds, while the number of Gaussian mixtures could be kept within a thousand (we used 512 Gaussian components in our experiments). This number is equivalent to that used for UBM in text-independent speaker verification. The difference here is that the Gaussian components are distributed among the tied states. Phone model is the obvious choice when the number of words is considerably large. Text-prompted speaker verification could be performed by constraining the vocabulary, for instance, to consist of digits from zero to nine. Under this scenario, whole word model is a better alternative. In addition to smaller number of models, word-based HMM could model the co-articulation effects within words better. Also, it is generally believed that speaker discriminant information presents mostly in such transitive regions of speech.

Given the speech recordings with orthographic transcriptions, HMM parameters are estimated by embedded training [12], where a graph of HMM models is composed according to the transcription. The forward-backward algorithm [13] is then applied on the graph to accumulate the sufficient statistics using the *soft* count, which means each frame can have non-zero posteriors for all the HMM states. Alternatively, in Viterbi training, the statistics are accumulated on the 1-best path. Each frame is aligned to a single HMM state with the posterior of 1. In our setup, Kaldi [14] is used for acoustic modeling which adopts Viterbi training.

2.2. Speaker dependent model

We train the speaker-independent (SI) model using the transcribed recordings from different speakers. The SI model thereby represents the phonetic and speaker variability that can be observed from the background set of speakers. From a Bayesian viewpoint, the parameters of the SI model serves as the prior knowledge where the speaker-dependent (SD) model could be derived. Speaker adaptation is applied to obtain the SD models given the enrollment utterances as depicted in Figure 1. In this way, both the lexical content and speaker characteristics are encoded in the SD model in the form of adapted HMMs.

There are several speaker adaptation techniques prevalent in the speech community, including the model-based MAP adaptation and the feature-based Constrained Maximum Likelihood Linear Regression (CMLLR) [15] adaptation. For this study, MAP is used since it is the most popular adaptation scheme for the text-independent speaker verification systems [2]. MAP aims to maximize the following posterior function:

$$\theta^{\text{MAP}} = \arg\max_\theta p(\theta|\mathbf{O}_1^T) = \arg\max_\theta p(\mathbf{O}_1^T|\theta)p(\theta) \quad (1)$$

where θ is the set of HMM parameters, while $\mathbf{O}_1^T$ denotes the feature vector sequence. Though the weights and covariance matrices could also be adapted, it is customary to adapt only the mean vectors:

$$\mu_{jm} = \frac{\tau_{jm}\mu^0_{jm} + \sum_{t=1}^T \gamma_{jm}(t)\mathbf{o}_t}{\tau_{jm} + \sum_{t=1}^T \gamma_{jm}(t)} \quad (2)$$

$$\gamma_{jm}(t) = \frac{p(\mathbf{o}_t|j,m)}{\sum_{m'=1}^M p(\mathbf{o}_t|j,m')} \quad (3)$$

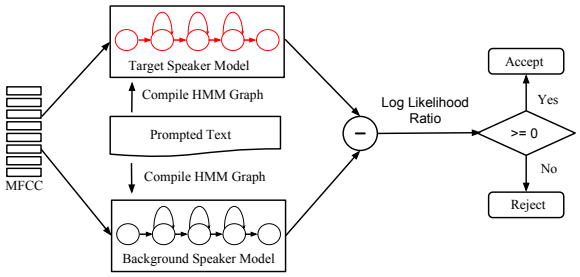

Figure 2: *HMM-based text-prompted speaker verification*

where μ^0_{jm} and Σ^0_{jm} are the mean and covariance of the SI model, μ_{jm} is the adapted mean for mixture m of state j, τ is a mixture-dependent smoothing constant. In this paper, τ is set to be a global constant. $\gamma_{jm}(t)$ is the posterior of mixture m of state j at time t. Note if the HMM state j is omitted, this reduces to the relevance MAP formulation for the GMM-UBM system used in the text-independent verification systems [16].

2.3. Text-prompted speaker verification

For text-prompted verification task, we consider four different trial types depending on whether the speaker is the target or not and whether the correct prompt is spoken or not. The target-correct trial corresponds to the scenario that the correct pass-phrase is spoken by the target speaker. The target-wrong includes the trials where the target speaker pronouncing the wrong pass-phrase. This will typically happen if an imposter plays a recorded speech segment of the target speaker. The imposter-correct trials are generated by the non-target speakers producing the prompted text. Lastly, the imposter-wrong trials are produced by the imposters speaking the wrong pass-phrases. The most straightforward approach is to use an ASR system to reject the wrong texts in addition to a TD speaker verification system. Besides the ASR training cost, the speakers may have different accents, native languages, ages, etc. Moreover, the verification environment may be very noisy. All these factors impose huge challenges for the ASR system. Therefore, we look into strategies to reject wrong-text trials using just the likelihood scores.

Figure 2 shows the flows of the speaker verification process using the HMM-based SD and SI models. Given the prompted text, a graph of HMM models is constructed according to a dictionary. In order to get the likelihood score of the test utterance given a model, Viterbi decoding is used to align the HMM graph against the input speech, yielding an alignment which labels each frame with a HMM state of the model. Once the alignment is available, the log likelihood of the utterance is computed as a summation of the log likelihoods of each HMM state along the alignment. Finally, a decision is made based on the log-likelihood-ratio between the target and background model.

3. Speaker and lexical model scoring

Based on the alignments generated by the background and target models, two sets of scores can be computed, of which the target scores comprises the contributions from both the speaker and lexical components. In order to identify the speaker, we need to suppress the scores associated with the lexical content. On the other hand, to reject the wrong texts, more emphasis should be put on the lexical components. To achieve this trade-off, various scoring schemes are proposed in this section.

3.1. Log-likelihood score computation

As mentioned above, the likelihood scores are computed from the alignments generated by the background and target models. Accordingly, two scoring schemes are studied depending on the

how the alignments are used:

Separate alignments Given an utterance, the target and background models are used to do the Viterbi decoding on the HMM graphs to get their respective alignments. This scoring approach has the advantage that each model will produce the best likelihood scores since their own alignments are used. However, there will be mismatches between the two alignments in terms of which HMM state a frame is aligned, adding more nuisance factors in the final log-likelihood-ratio scores. Given an alignment and the model parameter set, the log likelihood of a model given the utterance $\mathcal{X}$ is computed as:

$$\log P(\mathcal{X}|\theta^s) = \frac{1}{T_s}\sum_{t=1,q_t^s\neq \text{sil}}^{T} \log P(\mathbf{o}_t|\theta^s, q_t^s)$$
$$\log P(\mathcal{X}|\theta^u) = \frac{1}{T_u}\sum_{t=1,q_t^u\neq \text{sil}}^{T} \log P(\mathbf{o}_t|\theta^u, q_t^u)$$

where T_s and T_u are the number of voiced frames from the SD and SI alignments respectively, $\mathcal{X}$ represents the frame sequence of the utterance, $\log P(\mathbf{o}_t|\theta^s, q_t^s)$ is the log likelihood of the current observation $\mathbf{o}_t$ give the SD state q_t^s and θ^s, $\log P(\mathbf{o}_t|\theta^u, q_t^u)$ is the SI model likelihood given the SI parameter θ^u and the SI state q_t^u. Note that we have excluded the silence models for likelihood computation since the silence frames do not contain any speaker or lexical information.

Same alignment To circumvent the issues of HMM alignment mismatches, the same alignment, generated by either the target or the background model, can be used to compute the likelihoods for both models. Although one of the models will not produce its best score, scoring on the same alignment guarantees that each speech frame is scored against the same HMM model, where the background scores are computed from the SI HMM and the target scores are obtained from the same model with speaker adaptation. In this way, the lexical variabilities are suppressed in the final log-likelihood-ratio scores.

For scoring with separate alignments, the log-likelihood-ratio of the target and background model is formulated as:

$$\mathcal{L}_s(\mathcal{X}) = \log P(\mathcal{X}|\theta^s) - \log P(\mathcal{X}|\theta^u) \tag{4}$$

For the other scoring scheme, the ratio is computed using either the SD ($\mathcal{L}_s^s(\mathcal{X})$) or the SI ($\mathcal{L}_s^u(\mathcal{X})$) alignment:

$$\mathcal{L}_s^s(\mathcal{X}) = \frac{1}{T_s}\sum_{t=1,q_t^s\neq \text{sil}}^{T} (\log P(\mathbf{o}_t|\theta^s, q_t^s) - \log P(\mathbf{o}_t|\theta^u, q_t^s))$$
$$\mathcal{L}_s^u(\mathcal{X}) = \frac{1}{T_u}\sum_{t=1,q_t^u\neq \text{sil}}^{T} (\log P(\mathbf{o}_t|\theta^s, q_t^u) - \log P(\mathbf{o}_t|\theta^u, q_t^u))$$

3.2. Rejecting wrong text

The alignment-based scoring schemes in section 3.1 aim to remove the lexical variabilities by scoring the alignments from the same text. On the contrary, for the target-wrong and imposter-wrong trials, the main concern is to reject the wrong texts. To this end, this paper formulates the background scores as an interpolation of the SI decoding and the SI alignment scores. More specifically, the background score is computed as an average of the SI decoding and the SI alignment scores. The rational is that the SI decoding scores offer a measurement of the acoustic model confidence given the speech segment regardless of the text or the speaker. To compute the SI decoding score, a special word loop is used to decode the input speech by the background model. The obtained scores are purely generated by the acoustic model as the word loop has all zero language model scores.

For the trials with the wrong text, the decoding scores are usually larger than the alignment scores as the "wrong" texts are used for alignment. The average with the decoding scores will increase the SI scores thus reducing the likelihood ratio for the trials with the wrong texts. On the other hand, the effect of the average will be much lesser for the trials with the correct texts as the difference between the SI alignment score and decoding score is much smaller. Taking this into consideration, the log-likelihood-ratio for both schemes can now be computed as:

$$\mathcal{L}_s(\mathcal{X}) = \frac{1}{T_s}\sum_{t=1,q_t^s\neq \text{sil}}^{T} \log P(\mathbf{o}_t|\theta^s, q_t^s) \tag{5}$$
$$- \left[\frac{1}{T_u}\sum_{t=1,q_t^u\neq \text{sil}}^{T} \frac{\log P(\mathbf{o}_t|\theta^u, q_t^u)}{2} + \frac{1}{T_d}\sum_{t=1,q_t^u\neq \text{sil}}^{T} \frac{\log P(\mathbf{o}_t|\theta^u, q_t^d)}{2}\right]$$

where $\log P(\mathbf{o}_t|\theta^u, q_t^d)$ is the decoding score of state q_t^d given $\mathbf{o}_t$ on the best path and T_d is the number of voiced frames.

4. Digit-prompted speaker verification

4.1. Experimental setup

The RSR2015 [6] Part III background set is used to train the HMM models using Kaldi [14]. Since the corpus contains only ten digits from zero to nine, word-based HMMs are adopted. The features are the standard 20-dimensional MFCCs with its first and second derivatives yielding a dimension of 60. No voice activity detection (VAD) is applied. For the digit-prompted verification task, the three sequences of ten digits are used for enrollment and the ten sequences of five digits are used for verification. 72 out of 143 female speakers and 79 out of 157 male speakers are randomly chosen for this study. In addition, according to the protocol in [6], session {1,4,7}, are chosen for enrollment and the sentences from the remaining six sessions are for testing. For acoustic modeling, each digit is modeled as a left-to-right HMM with three emitting states and the total number of Gaussians is 512.

4.2. Speaker verification using different scoring schemes

We then report the EERs using the speaker-adapted models with different likelihood computation schemes in Table 1. The EERs are computed such that the genuine scores are provided by the target-correct trials and the non-target scores from the imposter-correct trials. In addition, the silence model scores and the transition probabilities are excluded during likelihood computation.

Table 1: *EER (%) comparison of scoring schemes*

	Separate Alignments	Rescore SI alignments	Rescore SD alignments
Female	3.82	2.51	3.60
Male	2.87	1.01	2.14

Rescoring SI (SD) alignments denotes that the sentence likelihood is computed on the alignments obtained from the SI (SD) model. Comparing the three scoring schemes, using the same alignment outperforms the separated alignment scheme significantly, especially rescoring on the SI alignments. By using the same alignments, both the target and background scores are computed with the same lexical constraint, i.e., the same HMM state for each frame. Therefore, the lexical variabilities are suppressed in the log-likelihood-ratio scores, leading to a better speaker verification performance.

Comparing the two rescoring methods, scoring on SI alignments outperforms scoring on SD alignments significantly. This can be explained by how the SD alignments are generated. For target-correct trials, the SD alignments are generated by the "true" speaker model, leading to better alignments. On the other hand, for the imposter-correct trials, the SD alignments are generated by the "imposter" models, resulting in poorer alignments.

The latter may play a major role since the number of imposter-correct trials is significantly larger than the target-correct trials. This may be the reason why the scoring on SI alignments performs better than the other two schemes.

Finally, if the silence models are not discarded, the EERs for the SI alignment scoring scheme are 2.21%/2.85% for the male and female genders respectively, a significant performance drop compared to the EER in Table 1. This clearly shows the importance of excluding the silence model scores for likelihood computation. The best EER by recoring SI alignments (1.01%/2.51%) is also significantly better than the best EERs (2.64% and 4.54%) reported in [9], where five JFA-based systems were fused. Therefore, in the following experiments, the likelihood scores are computed on the SI alignments.

5. The RedDots challenge

Finally, we evaluate the system under an unconstrained vocabulary scenario on the part 4 tasks of the RedDots challenge[1].

5.1. RedDots dataset

The RedDots [11] is a crowd-sourcing speech data collection initiative started in early 2015. The heterogeneous nature of the data is very challenging due to different handsets and the vast lexical variabilities in the recording. More importantly, most of the English speeches are spoken by nonnative English speakers. Since it is still an ongoing project, we use the first release which contains all the audited recordings up to August 17th 2015. It has 62 speakers including 49 male and 13 female speakers from 21 countries. The total number of sessions for the current release is 572 (473 male and 99 female sessions).

5.2. Experimental setup

For large vocabulary recognition task, finer acoustic modeling granularity is required since there will not be enough training data to train each word. To handle the co-articulation effects, triphone acoustic modeling are adopted [17, 18]. To model the vast lexical contents, WSJ0 (LDC93S6A) corpus together with the RSR2015 background set are used to train the phoneme-based HMM models. The phone set contains 40 phones including a silence model and each phone is modeled as a left-to-right HMM with three emitting states using Kaldi. The number of Gaussians is also fixed at 512 and 427 senones are obtained after decision tree state clustering [19].

5.3. Task 4: text-dependent enrollment

For the TD enrollment task, a speaker is enrolled by repeating one single sentence three times. The target model is represented as a pair of speaker and enrollment sentence. For verification, four trial types are considered, namely the target-correct (IC), target-wrong (TW), imposter-correct (IC), imposter-wrong (IW). The number of trials are given in Table 2. The EERs for each trial type are given in Table 3.

Table 2: Number of trials for the Reddots challenge part 4 task

	Target Correct	Imposter Correct	Target Wrong	Imposter Wrong
Female	1,122	3,906	25,806	180,462
Male	5,696	99,264	131,002	4,999,686

The likelihoods are computed on the SI alignments for both the target and the background models excluding the silence model scores. To reject the wrong texts, the background model scores are obtained by an average of the SI alignment scores and the SI decoding scores. To show the effectiveness of the SI decoding scores, a new set of EERs are computed using the log-likelihood-ratios between the target and background speaker models without the SI decoding scores. The new EERs are given in the last three columns of the same table. As we can see, without incorporating the SI decoding score, the EERs for the target-wrong and imposter-wrong trials are considerably worse.

Table 3: EERs on RedDots challenge part 4 TD enrollment task

	With Decoding Score			Without Decoding Score		
	IC	TW	IW	IC	TW	IW
Female	5.44	1.52	0.98	5.79	4.46	3.57
Male	3.67	1.62	1.19	3.79	4.25	2.99

5.4. Task 4: text-independent enrollment

In the TI enrollment task, a speaker is enrolled with free-text sentences. A total of ten sentences are used for the enrollment and the lexical content of the enrollment sentences are not the same. Each speaker is represented by the speaker ID only rather than a speaker-sentence pair. The EERs are given in table 4. From Table 4, significant EER degradation compared to the TD enrollment task is observed. This is expected since there are no lexical mismatches between enrollment and verification for the TD enrollment task. On the other hand, for the TI enrollment task, total lexical content mismatches exist between training and testing since the enrollment sentences do not appear in verification. The only information that the speaker model can reply on for the verification task is the speaker adapted senones. However, none of the speaker's enrollment data in the current release covers all the 40 phones. On average, each speaker's enrollment data cover only 34 phones. If the phones are not covered during enrollment, they will not be adapted. Consequently, the speaker model cannot distinguish them from the SI model, leading to the huge degradation in performance.

Table 4: EERs on RedDots challenge part 4 TI enrollment task

	Imposter Correct	Target Wrong	Imposter Wrong
Female	13.72	6.24	4.74
Male	11.6	7.38	6.57

6. Conclusion

We investigated the joint speaker and lexical modeling for text-prompted speaker verification given very short utterances. A HMM-based background model was trained to model the lexical space. Given the enrollment sentences, speaker adaptation was then applied to the background model to get the speaker-specific acoustic models to encode both the speaker and lexical characteristics. For verification, we developed a scoring model that scores only the speaker part rather than the joint speaker-lexical component leading to a better speaker verification performance. Experiments were conducted on the text-prompted task of RSR2015 and the RedDots datasets. For the RSR2015 task, our best single system gives an EER of 1.01% and 2.51% for male and female genders respectively using the target-correct and imposter-correct trials, a significantly improvement over the recently published JFA-based approach with fusion (2.65% and 4.54%). On the RedDots corpus, the joint speaker and lexical modeling system also yields very competitive results. Future work includes incorporating the DNN bottleneck features [20] to increase the robustness of the system against noise and channel factors.

[1] https://sites.google.com/site/thereddotsproject/

7. References

[1] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010.

[3] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," CRIM, Montreal, Tech. Rep., 2005.

[4] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Inter-Speaker Variability in Speaker Verification," *IEEE Trans. Audio, Speech, and Lang. Process.*, 2008.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[6] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, no. 0, pp. 56 – 77, 2014.

[7] L. Chen, K. Lee, B. Ma, W. Guo, H. Li, and L. Dai, "Phone-centric local variability vector for text-constrained speaker verification," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 229–233.

[8] N. Scheffer and Y. Lei, "Content matching for short duration speaker recognition," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 1317–1321.

[9] T. Stafylakis, P. Kenny, J. Alam, and M. Kockmann, "JFA for speaker recognition with random digit strings," in *INTERSPEECH*, 2015.

[10] J. luc Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[11] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brummer, D. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, J. Alam, A. Swart, and J. Perez, "The RedDots Data Collection for Speaker Recognition," in *INTERSPEECH*, 2015.

[12] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[13] L. R. Rabiner, "Readings in speech recognition," A. Waibel and K.-F. Lee, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, ch. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pp. 267–296. [Online]. Available: http://dl.acm.org/citation.cfm?id=108235.108253

[14] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The kaldi speech recognition toolkit," in *In IEEE ASRU 2011*, 2011.

[15] M. F. Gales, "The generation and use of regression class trees for MLLR adaptation," Cambridge University, Tech. Rep., 1996.

[16] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan. 2000.

[17] K.-F. Lee, "Readings in speech recognition," A. Waibel and K.-F. Lee, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, ch. Context-dependent Phonetic Hidden Markov Models for Speaker-independent Continuous Speech Recognition, pp. 347–366.

[18] G. Wang and K. C. Sim, "Regression-based context-dependent modeling of deep neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 11, pp. 1660–1669, Nov. 2014.

[19] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the Workshop on Human Language Technology*, ser. HLT 1994, 1994, pp. 307–312.

[20] F. Richardson, D. A. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letter*, vol. 22, no. 10, pp. 1671–1675, 2015.