Audio Engineering Society

Convention Paper 9844

Presented at the 143rd Convention
2017 October 18–21, New York, NY, USA

A Simplified 2-Layer Text-dependent Speaker Authentication System

Giacomo Valenti[1,2], Adrien Daniel[1], and Nicholas Evans[2]

[1]*NXP Software, Mougins, France*
[2]*EURECOM, Biot, France*

Correspondence should be addressed to Giacomo Valenti (giacomo.valenti@nxp.com)

ABSTRACT

This paper describes a variation of the well-known HiLAM approach to speaker authentication which enables reliable text-dependent speaker recognition with short-duration enrollment. The modifications introduced in this system eliminate the need for an intermediate text-independent speaker model. While the simplified system is admittedly a modest modification to the original work, it delivers comparable levels of automatic speaker verification performance while requiring 97% less speaker enrollment data. Such a significant reduction in enrollment data improves usability and supports speaker authentication for smart device and Internet of Things applications.

1 Introduction

The rapidly-growing smart device market and the explosion of the Internet of Things (IoT) has fueled the need for low footprint and efficient speaker authentication solutions, e.g. [1]. Unfortunately, many approaches to Automatic Speaker Verification (ASV) place unrealistic demands on enrollment and recognition/test data [2]. The need for anything more than a few seconds of speech impacts on usability and creates resistance among mass-market users.

ASV research has largely been driven by the Speaker Recognition Evaluations (SREs) administered by the US National Institute of Standards and Technology (NIST)[1]. These evaluations have typically focused on enrollment and testing with a duration in the order of a few minutes. While the SREs have stimulated tremendous progress over the last two decades, today's state-of-the-art speaker verification technology is often ill-suited to authentication applications which demand reliable recognition using utterances with a duration in the order of a few seconds [3, 4, 5]. With a clearly different use case scenario, the NIST SREs have also focused on text-independent recognition, whereas short-duration recognition generally calls for text-dependent operation.

State-of-the-art i-Vector and probabilistic linear discriminant analysis (PLDA) techniques are difficult to apply in text-dependent tasks [6, 7, 8] unless training data is plentiful [9] and unless impostor trials involve matching text [10]. Studies reported in [11, 12, 13, 14] demonstrated that joint factor analysis (JFA) systems can work well with little enrollment data, however, even under those conditions, both JFA and PLDA still rely on prior knowledge of the text content.

Initiatives dedicated to furthering progress in text-

[1] https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016

dependent recognition have gathered pace in recent years, prominent examples being the release of the RSR2015 [15] and RedDots [16] databases and associated evaluation campaigns. The RSR2015 database was furthermore introduced together with a baseline ASV system referred to as HiLAM (Hierarchical multi-Layer Acoustic Model) [6]. It involves a 3-layer approach to text-dependent speaker modeling. The HiLAM system is today a reference approach. Even it, though, is ill-suited to our target application since it learns an intermediate text-independent speaker model which in turn requires significant speaker enrollment data.

We have thus sought to develop an alternative to the HiLAM system which reduces demands on enrollment data for a short-duration, text-dependent speaker authentication application. Since the target application is text-dependent, the aim is to dispense with text-independent enrollment entirely. While an admittedly modest modification to the original work, the result is a simpler two-layer approach which achieves comparable ASV performance with a dramatic reduction in the need for enrollment data.

The remainder of this paper is organized as follows. Section 2 describes the RSR2015 database which was used for all experimental work reported herein. The original HiLAM baseline system is summarized in Section 3 whereas modifications to support short-duration speaker enrollment are presented in Section 4. A thorough comparison of the two systems performed using the standard RSR2015 evaluation protocol is presented in Section 5. Conclusions are presented in Section 6.

2 Database and Protocols

Almost all experimental work undertaken using the HiLAM system [6, 17, 18] is performed using the RSR2015 database [15]; the two were released almost in tandem and the database is distributed with protocols suited to the assessment of HiLAM-based text-dependent speaker verification systems. The RSR2015 database is one of the most versatile and comprehensive databases for such research. One particular aspect of RSR2015 which makes it better suited to this work than the more recent RedDots [19] successor is the particular speaker/part/session structure illustrated in Fig. 1. This is described below.

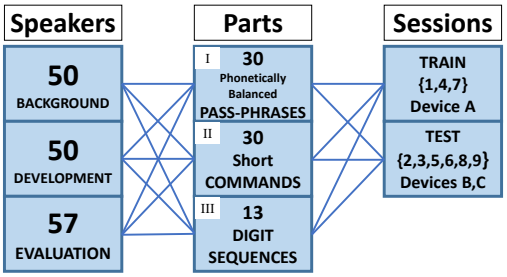

Fig. 1: *RSR2015 database partition for male speakers. The partition is identical for female speakers but with only 43 speakers in the evaluation set.*

2.1 Database

RSR2015 contains speech data collected from both male and female speakers and is partitioned into 3 evenly-sized subsets whose usual purpose is for background modeling, experimental development and evaluation. Each subset is comprised of 3 parts: phonetically-balanced sentences (part I), short commands (part II) and random digits (part III). Each part contains data collected in one of nine sessions. Three of these sessions are reserved for training while the remaining six are set aside for testing. The three training sessions are recorded using the same smart device (i.e. the same mobile phone or tablet) whereas the six testing sessions are recorded using two different smart devices.

Since our target application relates to short-duration pass-phrases, all experimental work reported in this paper was performed using part I data consisting of phonetically-balanced sentences. These are the same 30 Harvard sentences used in the collection of the better-known TIMIT database [6] which were designed to give a broad coverage of phonemes in the English language.

2.2 Training Protocol

Data reserved for background modeling is disjoint from training and testing data; there is no overlap in terms of speakers or sentences. Second-layer HiLAM models (GMMs) are trained with data from all three training sessions and all 30 sentences, totaling 90 utterances. Third-layer HiLAM models (HMMs) are trained with the three training utterances corresponding to each specific sentence (30 models each adapted from the second-layer model with three repetitions of each sentence).

Table 1: *The four different trial types used to assess the performance of a text-dependent speaker verification system. They involve different combinations of matching speakers and text.*

Trial Type	Speaker Match	Text Match
Target-Correct (TC)	Yes	Yes
Target-Wrong (TW)	Yes	No
Impostor-Correct (IC)	No	Yes
Impostor-Wrong (IW)	No	No

Initial experiments reported in this paper were performed using the standard protocols which are distributed with the RSR2015 database. However, since the goal of the work reported here is to reduce the quantity of data (number of utterances) needed for speaker enrollment, subsequent experiments were performed with subsampled versions of the standard protocols. As described later, the amount of data used for the learning of second-layer models is then either reduced (3-layer system with protocol sub-sampling) or eliminated entirely.

2.3 Testing Protocols

Test results reflect recognition performance estimated from a large number of single-utterance trials. Testing protocols used for all experiments are the standard part I testing protocols distributed with the RSR2015 database. All relate to one of the four trial types illustrated in Table 1. Any given trial involves either a target (model and test utterance correspond to the same speaker) or an impostor (model and test utterance correspond to different speakers). In addition, the text content either matches across model and test utterance (correct) or is different (wrong). This leads to three testing conditions which assess performance combining target-correct trials with trials of **one** mismatching combination: target-wrong, impostor-correct or impostor-wrong (note that target-wrong is therefore considered an impostor trial). The number of trials for each type in the standard RSR2015 protocols is illustrated in Table 2 for development and evaluation sets. The number of trials for each testing condition is TC+TW, TC+IC and TC+IW respectively. Finally, performance is expressed in terms of the equal error rate (EER).

Table 2: *Number of trials for Part I of the RSR2015 database for each of the four trial types illustrated in Table 1 and for development (Dev) and evaluation (Eval) subsets.*

Speaker-Text	Dev	Eval
Target-Correct (TC)	8,931	10,244
Target-Wrong (TW)	259,001	297,076
Impostor-Correct (IC)	437,631	573,664
Impostor-Wrong (IW)	6,342,019	8,318,132

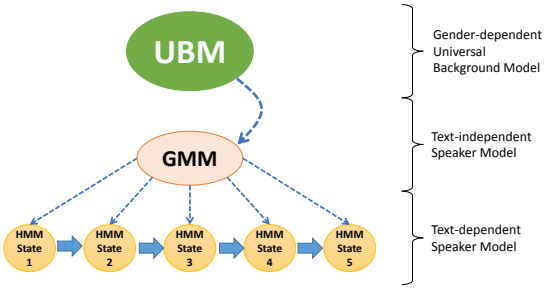

Fig. 2: *The original HiLAM system architecture reproduced from [17].*

3 The HiLAM Baseline

This section describes the original HiLAM architecture and essential elements of the basic algorithm. Maximum a posteriori (MAP) adaptation [20] is given particular attention; its optimization is fundamental to the simplified version of HiLAM presented later. Also presented are results for our specific implementation assessed using the RSR2015 database.

3.1 Architecture and Algorithm

The HiLAM system is a flexible, efficient and competitive approach to text-dependent automatic speaker verification. The architecture is illustrated in Fig. 2 and is composed of three distinct layers. They represent (i) a gender-dependent universal background model (UBM), (ii) a text-independent speaker model and (iii) a text-dependent speaker model. The first and second layers take the form of Gaussian mixture models (GMMs) whereas the third layer is a hidden Markov model (HMM).

The UBM is trained according to a conventional maximum likelihood / expectation maximization criterion [21]. The second layer text-independent speaker

model is derived from the UBM via MAP adaptation; this procedure is described in detail below. Different third-layer text-dependent speaker models are then learned for each sentence or pass-phrase. These take the form of 5-state, left-to-right HMMs. Each state of the HMM is initialized with the second layer text-independent GMM of the corresponding speaker and then adapted with several iterations of Viterbi realignment and retraining [22]. Each HMM therefore captures both speaker characteristics in addition to the time-sequence information which characterizes the sentence or pass-phrase. Full details of the HiLAM system in addition to the training and testing procedures can be found in [6].

3.2 MAP Adaptation

MAP adaptation is used to obtain the second-layer GMM from the first-layer UBM. A fundamental parameter of the MAP algorithm which governs the degree of adaptation is the so-called relevance factor, τ. Together with a probabilistic count of new data n_i for each Gaussian component i, it is used to determine an adaptation coefficient given by:

$$\alpha_i^\rho = \frac{n_i}{n_i + \tau^\rho} \tag{1}$$

where $\rho \in \{\omega, \mu, \sigma\}$ indicates the relevance factor for the weight, mean or variance parameters of the GMM. The adaptation coefficients are then used to obtain the new weight, mean and variance estimates according to:

$$\hat{\omega}_i = [\alpha_i^\omega n_i/T + (1-\alpha_i^\omega)\omega_i]\gamma \tag{2}$$

$$\hat{\mu}_i = \alpha_i^\mu E_i(x) + (1-\alpha_i^\mu)\mu_i \tag{3}$$

$$\hat{\sigma_i^2} = \alpha_i^\sigma E_i(x^2) + (1-\alpha_i^\sigma)(\sigma_i^2 + \mu_i^2) - \mu_i^2 \tag{4}$$

where each equation gives a new estimate from a combination of the respective training data posterior statistics with weight α and prior data with weight $(1-\alpha)$. T is a normalization factor for duration effects; γ is a scale factor which ensures the unity sum of weights. $E_i(x)$ and $E_i(x^2)$ are the first and second moments of posterior data whereas μ_i and σ_i^2 are the mean and variance of prior data, respectively [23].

In our experiments, each stage of adaptation is performed with a common value of τ, and hence α, for Equations 2, 3 and 4; the use of different values does not lead to better performance. Two distinct relevance factors are used at each stage, however: (i) for the adaptation of the UBM to the GMM, τ_1 and (ii) for the adaptation of the GMM to the HMM, τ_2. The first relevance factor, τ_1, acts to balance the contribution of the UBM and speaker-specific adaptation data to the parameters of the new speaker model, while the second, τ_2, controls adaptation between the text-independent and text-dependent speaker models.

3.3 Configuration and Performance

Silence removal is first applied to raw speech signals sampled at 16 kHz. This is performed according to ITU-T recommendation P.56[2] which specifies an active speech level of 15.9 dB. In practice this results in the removal of approximately 36% of the original data. The remaining 64% is then framed in blocks of 20ms with 10ms overlap. The feature extraction process is standard and results in 19 static Mel frequency cepstral coefficients (MFCC) without energy (C0). These are appended with delta and double-delta coefficients resulting in feature vectors of 57 dimensions.

The number of Gaussian components is empirically optimized. The literature shows that higher values (512-2048) are often used for text-independent tasks [24, 23] or with systems based on i-Vector and PLDA techniques [10, 25]. In contrast, lower values (128-256) are typically used in text-dependent tasks and techniques such as HiLAM [26, 17]. We obtained the best performance with 64 Gaussian components.

Results for our implementation of the HiLAM baseline are presented in Table 3 alongside those presented in the original work [27]. Results are presented for male speakers only and for the most challenging IC impostor condition. While results for our system are worse than those in the original work, performance is still respectable, with EERs of less than 2% for both development and evaluation subsets.

4 Simplified HiLAM

Described in this section are experiments which assess the necessity of text-independent enrollment and a number of modifications to the original HiLAM baseline system which enable competitive performance with

[2] http://www.itu.int/rec/T-REC-P.56-201112-I/en

Table 3: *Comparison of results for our implementation of the HiLAM system with original results reported in [27]. Results shown for male speakers in part I of the RSR2015 database and for the IC impostor condition.*

Subset	Our Implementation	Larcher et al. [27]
Development	1.63%	1.43%
Evaluation	1.81%	1.33%

greatly reduced durations of speaker enrollment data. Among these modifications is the reduction of the 3-layer approach to only two layers and associated re-optimization. The new system learns text-dependent speaker models using only three training utterances.

4.1 Enrollment Demands

The HiLAM system is well-suited to applications involving both text-independent *and* text-dependent speaker recognition scenarios. Satisfactory performance in these two scenarios calls for a large amount of training data; the original HiLAM system reported in [6] used 90 utterances for training middle layer text-independent speaker models.

The need for such an amount of enrollment data can be impractical, if not unusable in many cases such as smart device and Internet of Things applications. This paper extends the past work to address an exclusive text-dependent scenario which demands far less enrollment data. The following describes a simplified approach which eliminates the middle layer entirely and which delivers competitive text-dependent recognition with only three training utterances with only modest performance degradation.

4.2 Necessity of Text-Independent Enrollment

Our optimization of the original HiLAM system showed that the best performance is delivered with comparatively higher and lower values of τ_1 and τ_2 respectively (see Equation 1). This finding indicates that only modest adaptation is applied between layers 1 and 2, whereas more significant adaptation is applied between layers 2 and 3. This then calls into question the real need for text-independent enrollment or, in other words, the real need for the middle layer.

Table 4: *Performance for different durations of 2nd layer text-independent training. The last row shows results for the simplified HiLAM system with no text-independent training. Results shown for the RSR2015 development set and for the IC condition.*

	Number of utterances	EER
3-Layer	90+3	1.63%
	60+3	1.66%
	30+3	1.62%
	3	2.33%
2-Layer	3	1.84%

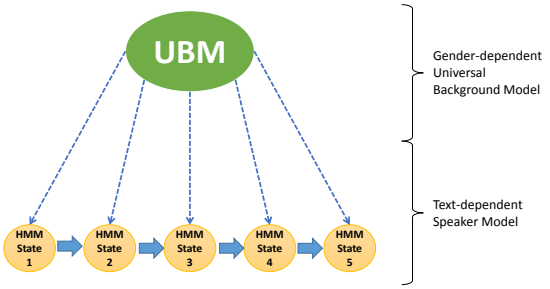

Fig. 3: *The simplified 2-layer architecture: text-dependent speaker models are adapted directly from the UBM.*

In order to assess the necessity of text-independent enrollment, we conducted a sequence of experiments in which the number of text-independent utterances used for layer-two training was successively subsampled from 90 to 60 and then 30 by taking 2 and 1 training sessions out of 3, respectively. Results are illustrated in Table 4. They show that performance remains unchanged as the quantity of text-independent enrollment data is reduced from 90 to 30 utterances. This finding suggests that text-independent enrollment may be unnecessary when the recognition task is ultimately text-dependent.

4.3 Layer Reduction

Given the observations reported above, we decided to assess performance when the middle layer, text-independent enrollment is dispensed with entirely. Speaker enrollment is then performed in text-dependent fashion exclusively as illustrated in Fig. 3. Each state

Table 5: *Comparison of results for the original work [27] and those obtained with the simplified system reported in this paper. Results for male speakers in part I of the RSR2015 database. (Results for each condition correspond to their combination with TC trials.)*

System	IC-Dev	TW-Dev	IW-Dev	IC-Eval	TW-Eval	IW-Eval
Larcher 3-Layer	1.43%	1.00%	0.20%	1.33%	0.66%	0.09%
Valenti 2-Layer	1.84%	1.09%	0.32%	1.24%	0.52%	0.05%

of the HMM speaker model is now initialized using the UBM instead of the speaker-specific text-independent GMM. Adaptation is otherwise the same as before and performed using the same three utterances of the same sentence. The number of Gaussian components (64) is left unchanged from the 3-layer implementation (see Section 3.3) and the single remaining relevance factor τ (3) is set to the same value of τ_2 (see Section 3.2). These parameters were found to be optimal in the case of the simplified system.

Results are illustrated in the last row of Table 4. Performance degrades slightly, from an EER of 1.6% for the baseline 3-layer system to 2.3% when enrollment is performed with only 3 speaker-specific utterances. Performance for the reduced 2-layer system improves slightly to 1.8%. Despite a reduction in enrollment data in the order of 97%, the increase in error rate is only 0.2%. Such a compromise between performance and usability would be quite acceptable in many practical scenarios.

5 Evaluation Results

Results presented above relate to the development set and the IC condition only. Presented in this section is a full performance comparison of the original HiLAM approach in [27] to the simpler 2-layer system presented in this paper using the full RSR2015 development and evaluation sets, including the three different test conditions, namely IC, TW and IW.

Results are illustrated in Table 5. The first row indicates the specific test condition for development (dev) and evaluation (eval) sets. Results presented in the original work [27] are illustrated in the second row whereas those for the new 2-layer system are presented in the third row. They correspond respectively to the full enrollment condition (90 text-independent utterances for layer 2 and 3 text-dependent utterances for layer 3) and the reduced enrollment condition (3 text-dependent utterances only). These results confirm the findings presented above, namely that significant improvements to usability can be delivered by reducing the demand for enrollment data with only modest increases in error rates. Both systems achieve better performance for the evaluation set than for the development set. While this finding is counter-intuitive, it is consistent with other results in the literature, e.g. [15, 17, 26, 27, 28], one possible explanation for which is differences in the distributions of recording devices across the two subsets.

Compared to the original work, performance for the 2-layer system deteriorates for the development set. In contrast, performance for the evaluation set improves. This result is particularly encouraging. The drop from 1.33% to 1.24% corresponds to a 7% relative reduction in the EER and comes with the same 97% reduction in demand for enrollment data. This is a significant improvement to usability in the case of text-dependent recognition.

6 Conclusions

This paper proposes a simplified version of the HiLAM approach to text-dependent automatic speaker verification in order to reduce the demand for speaker enrollment data. Many practical use case scenarios such as speaker authentication for smart device/home applications and those in the Internet of Things (IoT) domain call for enrollment with only a small number of passphrase repetitions. Experimental work presented in the paper questions the necessity of text-independent enrollment used in the conventional HiLAM system in the case that the ultimate recognition task is text-dependent in nature. Results produced using a publicly available, standard database and protocols show that text-independent, middle-layer enrollment impacts unnecessarily on usability. The paper shows that the middle layer of the HiLAM system and, hence, text-independent enrollment can be dispensed with entirely. Speaker enrollment is then performed using only three

repetitions of a given sentence or pass-phrase in a simplified two-layer approach. Since the collection of enrollment data is one of the most invasive and inconvenient tasks from the end user perspective, the usability of the new system improves greatly on the previous 3-layer HiLAM baseline system. The proposed approach, admittedly a modest modification of the original system, delivers largely comparable levels of automatic speaker verification performance with a 97% reduction in enrollment data.

References

[1] Lee, K. A., Ma, B., and Li, H., "Speaker Verification Makes Its Debut in Smartphone," in *IEEE SLTC Newsletter*, February 2013.

[2] Martınez, P. L. S., Fauve, B., Larcher, A., and Mason, J. S., "Speaker Verification Performance with Constrained Durations," in *International Workshop on Biometrics and Forensics (IWBF)*, IEEE, 2014.

[3] Kenny, P., Dehak, N., Ouellet, P., Gupta, V., and Dumouchel, P., "Development of the primary CRIM system for the NIST 2008 speaker recognition evaluation." in *INTERSPEECH*, pp. 1401–1404, 2008.

[4] Fauve, B. G., Evans, N. W., and Mason, J. S., "Improving the performance of text-independent short duration SVM-and GMM-based speaker verification," in *Odyssey Speaker and Language Recognition Workshop*, pp. 18–25, 2008.

[5] Poddar, A., Sahidullah, M., and Saha, G., "Performance comparison of speaker recognition systems in presence of duration variability," in *2015 Annual IEEE India Conference (INDICON)*, pp. 1–6, IEEE, 2015.

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

[7] Aronowitz, H., "Voice Biometrics for User Authentication," in *Afeka-AVIOS Speech Processing Conference 2012*, 2012.

[8] Sahidullah, M. and Kinnunen, T., "Local spectral variability features for speaker verification," *Digital Signal Processing*, 50, pp. 1–11, 2016.

[9] Stafylakis, T., Kenny, P., Ouellet, P., Perez, J., Kockmann, M., and Dumouchel, P., "I-Vector/PLDA variants for text-dependent speaker recognition," *CRIM Technical Report*, 2013.

[10] Stafylakis, T., Kenny, P., Ouellet, P., Perez, J., Kockmann, M., and Dumouchel, P., "Text-dependent speaker recognition using PLDA with uncertainty propagation," in *INTERSPEECH*, pp. 3651–3655, 2013.

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

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

[13] Kenny, P., Stafylakis, T., Alam, J., Ouellet, P., and Kockmann, M., "Joint Factor Analysis for Text-Dependent Speaker Verification," in *Odyssey Speaker and Language Recognition Workshop*, pp. 1705–1709, 2014.

[14] Stafylakis, T., Kenny, P., Alam, M. J., and Kockmann, M., "Speaker and channel factors in text-dependent speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24, pp. 65–78, 2016.

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

[16] Lee, K. A., Larcher, A., Wang, G., Kenny, P., Brummer, N., and others, "The RedDots data collection for speaker recognition," in *INTERSPEECH*, pp. 2996–3000, 2015.

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

[18] Larcher, A., Lee, K. A., Ma, B., and Li, H., "Imposture classification for text-dependent speaker

verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

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

[20] Lee, C.-H. and Gauvain, J.-L., "Speaker adaptation based on MAP estimation of HMM parameters," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pp. 558–561, IEEE, 1993.

[21] Bishop, C. M., *Pattern recognition and machine learning*, Information science and statistics, Springer, 2006.

[22] Rodríguez, L. J. and Torres, I., "Comparative study of the baum-welch and viterbi training algorithms applied to read and spontaneous speech recognition," in *Pattern Recognition and Image Analysis*, pp. 847–857, Springer, 2003.

[23] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, 10(1-3), pp. 19–41, 2000.

[24] Bimbot, F., Bonastre, J.-F., Fredouille, C., and others, "A tutorial on text-independent speaker verification," *EURASIP journal on applied signal processing*, 2004, pp. 430–451, 2004.

[25] Larcher, A., Bousquet, P.-M., Lee, K. A., Matrouf, D., Li, H., and Bonastre, J.-F., "I-vectors in the context of phonetically-constrained short utterances for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4773–4776, IEEE, 2012.

[26] Larcher, A., Bonastre, J.-F., and Mason, J., "Reinforced temporal structure information for embedded utterance-based speaker recognition." in *INTERSPEECH*, pp. 371–374, 2008.

[27] Larcher, A., Lee, K. A., Martınez, P. L. S., Nguyen, T. H., Ma, B., and Li, H., "Extended RSR2015 for text-dependent speaker verification over VHF channel," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[28] Larcher, A., Lee, K. A., Ma, B., and Li, H., "Phonetically-constrained PLDA modeling for text-dependent speaker verification with multiple short utterances," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7673–7677, IEEE, 2013.