INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Unified Framework for Speaker and Utterance Verification

Tianchi Liu, Maulik Madhavi, Rohan Kumar Das, Haizhou Li

Department of Electrical and Computer Engineering,
National University of Singapore, Singapore

liutianchi@u.nus.edu, {maulik.madhavi, rohankd, haizhou.li}@nus.edu.sg

Abstract

Speaker and utterance verification are two tasks that co-exist in text-dependent speaker verification (SV), where a phrase of the same lexical information is spoken during train and test sessions. The conventional approaches mostly verify the speaker and the utterance separately using two models. While there are studies on joint modeling of speaker and utterance, it is always desirable to have a common framework that performs both speaker and utterance verification to access the intended service. To this end, we propose a unified framework that deals with both objectives and the trade-off between the two. The unified framework is based on long short term memory network trained using both speaker and utterance information. We use Part I of RSR2015 database for the studies in this work. We show that the unified framework not only demonstrates competitive SV performance, but also provides a solution for a system to address different levels of security need.

Index Terms: text-dependent, speaker verification, utterance verification, unified framework, RSR2015

1. Introduction

Speaker verification (SV) aims to verify the claimed identity of a person for a given speech [1]. There are two broad categories of SV, namely, text-dependent and text-independent based on the nature of speech to be captured during train and test sessions [2, 3]. The former uses phases of same lexical content during training and testing. On the other hand, the latter does not need any restriction on lexical content for speaker modeling as well as testing. Hence, it requires a larger amount of speech data for modeling speaker characteristics. Studies show that performance of such system degrades significantly as train/test data reduce [4, 5]. As a result, text-dependent SV is more practical in many voice biometrics applications for access control [6, 7].

The classical approaches for text-dependent SV utilize dynamic time warping to compare the speech [8]. In the recent years, standard databases such as RSR2015 and RedDots provide a common platform for technology development [9, 10]. A hybrid model, namely hierarchical multi-layer acoustic model (HiLAM) based on Gaussian mixture model (GMM) and hidden Markov model (HMM) is proposed that utilizes the sequence as well as speaker information [9]. It has been shown that HiLAM outperforms i-vector based speaker modeling that is widely popular in text-independent SV [9, 11]. Few other novel approaches in text-dependent speaker modeling include joint factor analysis [12], unsupervised HMM-universal background model (UBM) [13], i-vector/HMM [14] and j-vector [15]. Additionally, various deep learning techniques have been proposed for fixed phrase SV in recent times [16–23].

In text-dependent SV, utterance verification can be seen as a subtask. A comparison of different features and modeling techniques for text-dependent SV can be found in [24]. Joint speaker and utterance models with HMM triphone models are used in [25]. The phonetic posteriorgrams derived from GMM and deep neural network (DNN) frameworks have been found to be useful to capture the lexical information for text-dependent SV [26, 27]. Similarly, benefit of DNN based speaker embeddings with content information is shown in [28]. On the other side, few works focus towards compensating the lexical content information to improve the SV performance [29–31]. These works altogether depict the importance of content modeling as well as normalization for fixed phrase SV. However, when we need to evaluate the utterance verification as a task independently, the joint modeling of speaker and lexical information may not be the best way. Therefore, most of the works in the literature focus on individual systems to perform utterance verification [32].

The utterance verification has been explored previously in the context of continuous speech recognition [33, 34]. It has not been explored much in the context of text-dependent SV. The general way of utterance verification in the context of text-dependent SV is to have a separate framework from the SV system [32, 35]. Hence, a single system which is capable of performing both speaker and utterance verification is more desirable in practice. From engineering point of view, speech recognition and speaker recognition are independent tasks. However, human brain interprets and decodes the information from speaker traits and linguistic content from the speech in joint corroborative manner [36, 37]. Similarly, a unified framework for speaker and language recognition has been attempted using shared DNN that outperforms the single task implementation [38]. These studies motivated us to explore a unified framework that performs both speaker and utterance verification to use in multiple application scenarios.

In this work, we propose a unified framework that can perform both speaker and utterance verification together. We refer this unified framework as speaker-utterance-verification (SUV) framework from here on. The system is developed using stack of long short term memory networks (LSTM) that learns both speaker and utterance models from the training data. In particular, we propose to build a shared LSTM to drive another two separate LSTM layers performing specific task of speaker and utterance verification. Thus, there are two outputs from the system in terms of speaker and utterance posteriors that can be combined according to the requirement of speaker or utterance verification. The studies are performed on RSR2015 database for the fixed phrase scenarios.

The remainder of the paper is organized as follows. Section 2 details the proposed unified framework for speaker and utterance verification. In Section 3, we present the system description. Section 4 reports the results and discussion. Finally, the work is concluded in Section 5.

http://dx.doi.org/10.21437/Interspeech.2019-1994

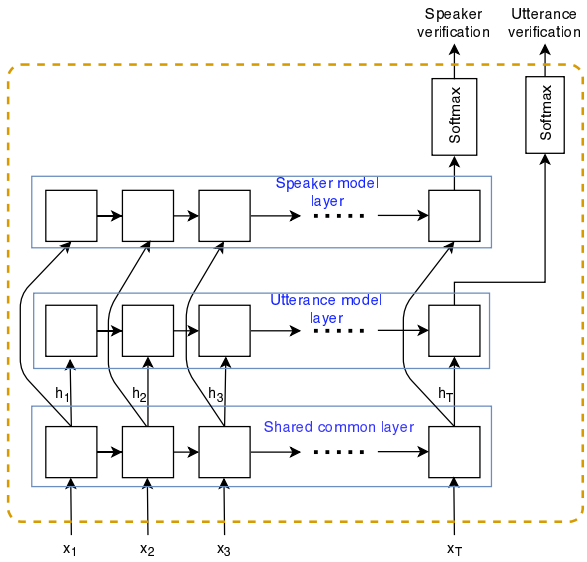

Figure 1: *Training phase for proposed unified SUV framework.*

2. Unified SUV Framework

This section describes the unified SUV framework to model speaker and utterance information. In this framework, we use both speaker and utterance information to train both speaker and utterance models, respectively. The DNN and LSTM based recurrent neural network are two possible ways of such modeling [36, 39]. The former supports the input to have fixed length, whereas recurrent architecture allows different length input. Our unified SUV framework has recurrent architecture to characterize the temporal dynamics during modeling.

The training phase of proposed unified SUV framework for speaker and utterance verification is shown in Figure 1. As observed in the earlier studies, the strong correlation between speech and speaker recognition inspired us to use the shared common LSTM layer that acts as a front-end for models. The output from the shared layer is input to the both models, which consists of LSTM layers. The last output from each layer is given to fully connected layer to produce the speaker and utterance verification scores. The algorithm for training of the proposed unified SUV framework is described in Algorithm 1. The implementation of SUV framework is made available[1] for use. Next, we discuss the unified SUV framework in relation to the existing text-dependent systems and security trade-off associated with it.

2.1. Differences from Existing Frameworks

- We consider the end-to-end framework [21,36] as a reference, which is optimized only for SV task.
- Our proposed unified SUV framework employs a recurrent network architecture to encode the temporal dynamics of text dependent phrases, whereas the end-to-end network in [20] and the joint framework in [22] use DNNs without recurrent architecture. The absence of recurrent nature may not be helpful to capture and characterize temporal dynamics.
- The study in [22] considered joint training of j-vector and Bayesian model for SV studies. Later, it uses the Siamese network to extract j-vectors and the joint Bayesian model backends for scoring. However, the unified SUV framework is simple as it does not require additional models for scoring, such as Bayesian model in comparison to different versions of j-vector systems and yet serves for two tasks.
- The study presented in [25] uses an automatic speech recognition to jointly model speaker and lexical information GMM/HMM framework. However, we do not use any such additional speech recognition system and GMM/HMM for speaker and utterance verification in this work.
- The study in [31] normalizes the utterance content information from joint speaker and utterance model. As it compensates the lexical information to improve SV performance, it is not a suitable approach for utterance verification.

[1] https://github.com/sn1ff1918/SUV

Algorithm 1 Training algorithm for the proposed unified SUV framework

Input: Input feature sequence $X := \{\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_T\}$, speaker label y_{spk} and utterance label y_{utt}. (Note: In implementation, we used the batch of feature sequence instead of single sequence)

Output: Unified speaker-utterance model for speaker and utterance prediction.

1: **while** each epoch i **do**
2: Feed input sequence to shared common LSTM. $H = \mathrm{LSTM}_{comm}(X)$
where, H is the sequence of hidden vectors, i.e., $H := \{\mathrm{h}_1, \mathrm{h}_2, \cdots, \mathrm{h}_T\}$
3: Feed output of shared common LSTM to utterance model LSTM layer. $out_{utt} = \mathrm{LSTM}_{utt}(H)$
4: Feed output of shared common LSTM to speaker model LSTM layer. $out_{spk} = \mathrm{LSTM}_{spk}(H)$
5: Use fully-connected layers to predict the labels $p_{utt} = softmax(out_{utt}[T])$ and $p_{spk} = softmax(out_{spk}[T])$
6: Loss computation: $L := loss_{tot} = loss_{spk} + loss_{utt}$. where $loss_{spk}$ and $loss_{utt}$ indicate the categorical cross-entropy loss with respect to corresponding speaker and utterance model. To compute this, use predictions p_{spk}, p_{utt}, and speaker label y_{spk} and utterance label y_{utt}
7: Do back-propagation to compute the weights $\frac{\partial loss_{utt}}{\partial w_{utt}}$ and $\frac{\partial loss_{spk}}{\partial w_{spk}}$
8: Stop when converges or after fixed number of epochs
9: **end while**

2.2. Security Trade-off

Most of the text-dependent SV studies are focused on the overall SV performance. However, there may be different levels of security associated with respect to various application scenarios. For example, one may favor utterance verification over SV for low security environments, or favor SV in high security platforms. The proposed unified SUV framework leverages the administrator to implement such a trade-off according to the associated security with the intended application service. As one system single can perform both the tasks in the SUV framework, it is adaptable to multiple application scenarios.

3. System Description

In this section, the details of the SV system developed in this work are mentioned. The database, front-end processing and experimental setup are described in the following subsections.

Table 1: *A summary of RSR2015 corpus.*

Subset	# Speakers	
	Male	Female
Background	50	47
Development	50	47
Evaluation	57	49

3.1. Database

We conduct the studies on RSR2015 corpus [9]. The database comprises of 300 speakers data from 143 female and 157 male speakers. There are three different parts of the database depending on the nature of the fixed phrases used. The Part I contains 30 fixed phrase utterances of 3-4 seconds duration. Similarly, there are 30 fixed short commands of 1-2 seconds duration in Part II. The random digit based fixed sequences are kept in Part III of the corpus. Every phrase is spoken for 9 sessions by all the speakers, in which first, fourth and seventh sessions are considered for modeling the speakers. The rest of the sessions are used for testing.

Additionally, the corpus is organized in three sets, which are background, development and evaluation set. The 300 speakers of the corpus are divided uniformly across all these three sets. The detailed composition of the speakers can be observed from Table 1. The corpus follows four categories of trials, which are *Target Correct* (TC), *Impostor Correct* (IC), *Target Wrong* (TW) and *Impostor Wrong* (IW). There are three different test conditions to report performance according to the protocol mentioned in [9]: *Impostor Correct* where Target and impostor users speak the same pass phrase, *Target Wrong* where target users speak different pass phrases, and *Impostor Wrong* where target and impostor users speak different pass phrases. Equal error rate (EER) is used as the metric for evaluating the performance. In this work, we have considered Part I of the database for this study.

3.2. Front-end Processing

We consider short-term processing on the speech utterances with 20 ms frame size and 10 ms shift to extract 60-dimensional (20-base + 20-Δ + 20-$\Delta\Delta$) mel frequency cepstral coefficient (MFCC) features using KALDI[1] toolkit. The features are then normalized in the cepstral domain with cepstral mean and variance normalization (CMVN) using utterance level mean and variance statistics [8].

3.3. Experimental Setup

For our studies, the network takes 60-dimensional MFCC feature vectors. The network comprises of three LSTM layers as described earlier in Section 2. The shared common layer takes MFCC features as input and the output is distributed to both speaker and utterance modeling layers. The dimension of hidden layer is 256 in each case. The batch size of 128 is taken for all experiments. We kept learning rate as 0.01 with stochastic gradient descent optimizer. The training is done using PyTorch[2] toolkit. Table 2 shows the total number of trials in Part I of RSR2015 database. In order to produce the scores with respect to all four possible categories of trials, we adopted the strategies shown in Figure 2. In particular, the scores from speaker and utterance verification models for i^{th} speaker and j^{th} utterance are

[1]http://kaldi-asr.org/

[2]https://pytorch.org/

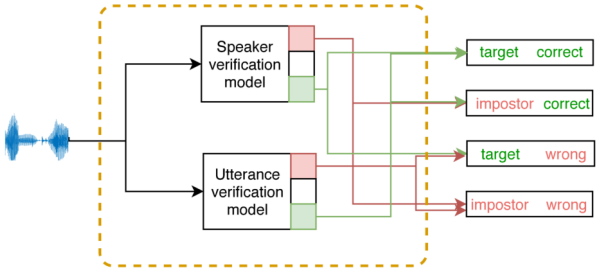

Figure 2: *Score computation for the four trial categories following the RSR2015 corpus test conditions.*

Table 2: *Number of trials for Part I of the RSR2015 database.*

Trial Categories	Male		Female	
	Dev	Eval	Dev	Eval
TC	8,931	10,244	8,419	8,631
TW	259,001	297,076	244,123	250,299
IC	437,631	573,664	387,230	414,249
IW	6,342,019	8,318,132	5,612,176	6,006,596

combined as follows:

$$S(spk^i, utt^j) = \log(p^{(i)}_{spk}) + \log(p^{(j)}_{utt}) \quad (1)$$

For a given test utterance during testing, all scores are computed with respect to speaker and utterance verification models. Further, as shown in Figure 2 to produce the score for all four categories, we pooled the prediction score from individual model. As discussed earlier, there are three different test conditions, namely, IC, TW and IW.

As a common reference, we consider the HiLAM and standard i-vector systems described in [9] for comparison. In addition, we consider the joint speaker-utterance system that is trained using speaker-utterance model by UBM adaptation. The joint speaker-utterance model system is slightly different than HiLAM. In HiLAM system, UBM data is first adapted to all utterances from the target speaker and then adapted to particular utterance. On the contrary, joint speaker-utterance model system directly adapts the UBM to particular speaker-utterance pair. Further, the joint speaker-utterance framework implemented here does not use the lexical information to train the triphone HMM and additional re-scoring model as used in previous studies [25, 31]. We also compare our results with different versions of recent j-vector systems reported in [22].

4. Results and Discussion

We now report the experimental results with detailed analysis. As discussed, we considered three systems as primary reference systems for the comparison, namely, HiLAM, i-vector, and joint speaker-utterance model systems. Among these systems, the i-vector system does not consider the temporal dynamics modeling. The results of different systems are shown in Table 3. We observe that the joint-speaker utterance model gives relatively better performance for IC trial category than HiLAM system. The HiLAM system does two level of adaptation, first at speaker level and then at speaker-utterance level that improves speaker discrimination. As shown in Table 3, the unified SUV framework outperforms these reference systems in most of the cases. The results validate the proposal of unified SUV framework for both speaker and utterance verification.

Table 3: *Performance in EER (%) for proposed unified SUV framework with comparison to existing frameworks on RSR2015 corpus.*

System	Male						Female					
	Dev			Eval			Dev			Eval		
	TW	IC	IW	TW	IC	IW	TW	IC	IW	TW	IC	IW
i-vector [9]	2.870	5.950	0.740	1.950	4.030	0.320	3.050	7.870	0.940	1.910	6.610	0.750
HiLAM [9]	1.660	3.690	0.490	0.820	2.470	0.190	1.770	3.240	0.450	0.610	2.960	0.140
Joint speaker-utterance	5.565	1.981	1.792	5.125	2.079	0.888	5.179	1.699	0.831	3.110	1.453	0.499
Unified SUV	**0.470**	**1.590**	**0.101**	**0.293**	**1.757**	**0.039**	**1.176**	4.323	**0.178**	**0.375**	2.009	**0.068**

Table 4: *Performance in EER (%) of different systems on the evaluation set of RSR2015 Part I. Here, J :j-vector with cosine similarity, JB :j-vector system with joint Bayesian model, J2 :joint training of j-vector extractor and joint Bayesian and the Siamese network for j-vector extractor and the joint Bayesian as a back-end, and J3 : joint training of j-vector extractor and joint Bayesian, and use the Siamese network output for verification.*

EER (%)	J	JB	J2	J3	Unified SUV
IW	0.95	0.02	0.02	0.02	0.06
TW	3.14	0.03	0.02	0.02	0.46
IC	7.86	3.61	2.81	2.42	**2.41**

Table 5: *Performance in EER (%) for speaker and utterance verification (UV) using unified SUV framework on Part I evaluation set of RSR2015 database.*

Tasks	Male	Female
SV	1.796	1.918
UV	0.021	0.011

Next, we compare our results with other deep learning systems. We combine male and female data as in [16, 17, 22] for fair comparison. Table 4 shows performance comparison of the proposed framework with other deep learning systems for the three test conditions. Most of these other systems are developed on the joint vector (i.e., j-vector) approach that uses a multi-task learning to extract hidden embedding representation. The results for j-vectors are cited from [22]. Table 4 shows that joint Bayesian back-end and Siamese network for scoring improve the performance of j-vectors. In addition, our proposed unified SUV framework gives better performance for IC trial category, whereas it is comparable in IW and perform poorer in TW trial category. Overall, unified SUV framework is comparable to other deep learning systems and at the same time easily adjustable to address different level of security need. In addition, the advantage of unified SUV framework compared to the j-vector based systems is that it does not use any additional Bayesian model. We also note that the unified SUV framework is suitable for a closed set scenario.

We then analyze the performance for SV and utterance verification separately. To evaluate SV system, we pooled all scores belonging to *Target* and *Impostor* categories. Similarly, we pooled all scores belonging to *Correct* and *Wrong* categories to evaluate the performance for utterance verification. We note that for utterance verification studies, we do not have a baseline as most of the works in the literature targets only for SV. Table 5 shows the performance for both speaker and utterance verification. The high performance achieved for both speaker and utterance verification shows the importance of the unified SUV framework to handle both tasks.

Furthermore, we introduce a weight factor in score computation as a trade-off for the unified SUV to work in high and low security scenarios, respectively, as discussed in Section 2. The

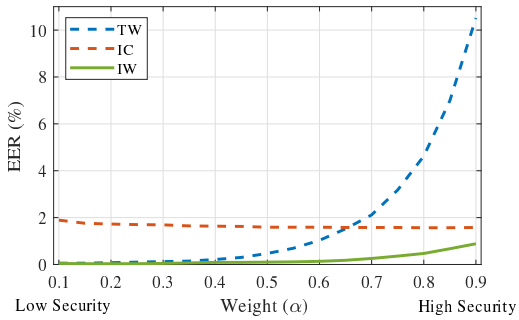

Figure 3: *Analysis of unified SUV framework for speaker and utterance verification (UV) on development set Part I for male.*

weight factor adjusts the relative bias towards particular task. To incorporate this, we modify Equation (1) as follows:

$$S(spk^i, utt^j) = \alpha \log(p^{(i)}_{spk}) + (1-\alpha)\log(p^{(j)}_{utt}) \quad (2)$$

where α is the relative weight that adjusts the balance between speaker and utterance model. Figure 3 shows the plots for three different test conditions with respect to weight (α) on Part I development set of male speakers. It can be observed that lower EER for TW trial category around lower value of α adjusts the unified SUV framework for better utterance verification. Similarly, as weight α increases that improves the SV performance as observed in IC trial category. Thus, the security trade-off between high and low scenarios can be adjusted in the proposed unified SUV framework as depicted in Figure 3.

5. Conclusions

This work presents a unified SUV framework for both speaker and utterance verification with a single system using recurrent LSTM. We used the stack of LSTM layers to model individual tasks. As compared to other deep learning approaches, this approach directly produces the scores for speaker and utterance verification. We evaluated the unified SUV framework on all the three test conditions for male and female speakers using RSR2015 Part I corpus. The result validates our unified SUV proposal. Further, the proposed unified SUV framework has the leverage to deal with speaker and utterance verification tasks that is performed by a weighted approach. This showcases its potential for high as well as low security application scenarios.

6. Acknowledgment

This research is supported by Programmatic Grant No. A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain).

7. References

[1] J. P. Campbell, "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] M. Hèbert, "Text-dependent speaker recognition," *Springer-Verlag Heidelberg*, pp. 743–762, 2008.

[3] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, pp. 12 – 40, 2010.

[4] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, , and M. Mason, "i-vector based speaker recognition on short utterances," in *Interspeech*, Florence, Italy, 2011, pp. 2341–2344.

[5] R. K. Das and S. R. M. Prasanna, *Speaker Verification for Variable Duration Segments and the Effect of Session Variability*. Lecture Notes in Electrical Engineering: Springer, 2015, ch. 16, pp. 193–200.

[6] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities," *IET Biometrics*, vol. 7, no. 2, pp. 91–101, 2018.

[7] R. K. Das and S. R. M. Prasanna, "Speaker verification from short utterance perspective: A review," *IETE Technical Review*, vol. 35, no. 6, pp. 599–617, 2018.

[8] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr 1981.

[9] 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.

[10] K. A. Lee, A. Larcher, W. Guangsen, K. Patrick, 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*, Dresden, Germany, 2015, pp. 2996–3000.

[11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Acoust., Speech & Signal Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[12] P. Kenny, T. Stafylakis, J. Alam, P. Ouellet, and M. Kockmann, "Joint factor analysis for text-dependent speaker verification," in *Odyssey*, Joensuu, Finland, 2014, pp. 200–207.

[13] A. K. Sarkar and Z.-H. Tan, "Text dependent speaker verification using un-supervised HMM-UBM and temporal GMM-UBM," in *Interspeech*, San Francisco, 2016, pp. 425–429.

[14] H. Zeinali, H. Sameti, L. Burget, J. Černock, N. Maghsoodi, and P. Matjka, "i-vector/HMM based text-dependent speaker verification system for RedDots challenge," in *Interspeech*, San Francisco, 2016, pp. 440–444.

[15] N. Chen, Y. Qian, and K. Yu, "Multi-task learning for text-dependent speaker verification," in *Interspeech*, Dresden, Germany, 2015, pp. 185–189.

[16] 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.

[17] Z. Shi, L. Liu, M. Wang, and R. Liu, "Multi-view (joint) probability linear discrimination analysis for j-vector based text dependent speaker verification," in *ASRU*, Okinawa, Japan, 2017, pp. 614–620.

[18] G. Bhattacharya, P. Kenny, J. Alam, and T. Stafylakis, "Deep neural network based text-dependent speaker verification : Preliminary results," in *Odyssey 2016*, Bilbao, Spain, 2016, pp. 9–15.

[19] H. Zeinali, H. Sameti, L. Burget, and J. Cernocký, "Text-dependent speaker verification based on i-vectors, neural networks and hidden markov models," *Computer Speech & Language*, vol. 46, pp. 53–71, 2017.

[20] H. Heo, J. Jung, I. Yang, S. Yoon, and H. Yu, "Joint training of expanded end-to-end DNN for text-dependent speaker verification," in *Interspeech*, Stockholm, Sweden, 2017, pp. 1532–1536.

[21] S. Dey, S. Madikeri, and P. Motlicek, "End-to-end text-dependent speaker verification using novel distance measures," in *Interspeech*, Hyderabad, India, 2018, pp. 3598–3602.

[22] Z. Shi, L. Liu, H. Lin, and R. Liu, "Joint learning of j-vector extractor and joint bayesian model for text dependent speaker verification," in *Interspeech*, Hyderabad, India, 2018, pp. 1076–1080.

[23] Z. Shi, M. Wang, L. Liu, H. Lin, and R. Liu, "A double joint bayesian approach for j-vector based text-dependent speaker verification," in *Odyssey*, France, 2018, pp. 365–371.

[24] Y. Liu, L. He, Y. Tian, Z. Chen, J. Liu, and M. T. Johnson, "Comparison of multiple features and modeling methods for text-dependent speaker verification," in *ASRU*, Okinawa, Japan, 2017, pp. 629–636.

[25] G. Wang, K. A. Lee, T. H. Nguyen, H. Sun, and B. Ma, "Joint speaker and lexical modeling for short-term characterization of speaker," in *Interspeech*, San Francisco, 2016, pp. 415–419.

[26] S. Jelil, R. K. Das, R. Sinha, and S. R. M. Prasanna, "Speaker verification using Gaussian posteriorgrams on fixed phrase short utterances," in *Interspeech*, Dresden, Germany, 2015, pp. 1042–1046.

[27] S. Dey, S. Madikeri, M. Ferras, and P. Motlicek, "Deep neural network based posteriors for text-dependent speaker verification," in *ICASSP*, Shanghai, China, 2016, pp. 5050–5054.

[28] S. Dey, T. Koshinaka, P. Motlicek, and S. Madikeri, "DNN based speaker embedding using content information for text-dependent speaker verification," in *ICASSP*, Calgary, Alberta, Canada, 2018, pp. 5344–5348.

[29] N. Scheffer and Y. Lei, "Content matching for short duration speaker recognition," in *Interspeech*, Singapore, 2014, pp. 1317–1321.

[30] S. Dey, S. Madikeri, P. Motlicek, and M. Ferras, "Content normalization for text-dependent speaker verification," in *Interspeech*, Stockholm, Sweden, 2017, pp. 1482–1486.

[31] R. K. Das, M. Madhavi, and H. Li, "Compensating utterance information in fixed phrase speaker verification," in *APSIPA ASC*, Hawaii, USA, 2018, pp. 1708–1712.

[32] T. Kinnunen, M. Sahidullah, I. Kukanov, H. Delgado, M. Todisco, A. K. Sarkar, N. B. Thomsen, V. Hautamki, N. Evans, and Z.-H. Tan, "Utterance verification for text-dependent speaker recognition: A comparative assessment using the RedDots corpus," in *Interspeech*, San Francisco, 2016, pp. 430–434.

[33] M. G. Rahim, C.-H. Lee, and B.-H. Juang, "A study on robust utterance verification for connected digits recognition," *The Journal of the Acoustical Society of America*, vol. 101, no. 5, pp. 2892–2902, 1997.

[34] E. Lleida and R. C. Rose, "Utterance verification in continuous speech recognition: decoding and training procedures," *IEEE Trans. on Acoust., Speech & Audio Process.*, vol. 8, no. 2, pp. 126–139, March 2000.

[35] H. Zeinali, L. Burget, H. Sameti, and H. Cernocky, "Spoken pass-phrase verification in the i-vector space," in *Odyssey*, France, 2018, pp. 372–377.

[36] Z. Tang, L. Li, D. Wang, and R. Vipperla, "Collaborative joint training with multitask recurrent model for speech and speaker recognition," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 25, no. 3, pp. 493–504, 2017.

[37] R. Kumar, V. Yeruva, and S. Ganapathy, "On convolutional LSTM modeling for joint wake-word detection and text dependent speaker verification," in *Interspeech*, Hyderabad, India, 2018, pp. 1121–1125.

[38] F. Richardson, D. A. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," in *Interspeech*, Dresden, Germany, 2015, pp. 1146–1150.

[39] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *ICASSP*, Shanghai, China, 2016, pp. 5115–5119.