Speaker-Utterance Dual Attention for Speaker and Utterance Verification

Tianchi Liu[1], Rohan Kumar Das[2], Maulik Madhavi[2], Shengmei Shen[1], Haizhou Li[2]

[1]Pensees Pte Ltd, Singapore
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore

{liutianchi, jane.shen}@pensees.ai, {rohankd, maulik.madhavi, haizhou.li}@nus.edu.sg

Abstract

In this paper, we study a novel technique that exploits the interaction between speaker traits and linguistic content to improve both speaker verification and utterance verification performance. We implement an idea of speaker-utterance dual attention (SUDA) in a unified neural network. The dual attention refers to an attention mechanism for the two tasks of speaker and utterance verification. The proposed SUDA features an attention mask mechanism to learn the interaction between the speaker and utterance information streams. This helps to focus only on the required information for respective task by masking the irrelevant counterparts. The studies conducted on RSR2015 corpus confirm that the proposed SUDA outperforms the framework without attention mask as well as several competitive systems for both speaker and utterance verification.

Index Terms: text-dependent speaker verification, utterance verification, attention, masking, RSR2015

1. Introduction

Speaker verification (SV) aims to verify the claimed identity of a person using given speech [1]. Its implementation is broadly categorized into text-dependent and text-independent based on the spoken contents used for enrollment and testing [1]. The former deals with use of fixed short phases, while the latter doesn't put any constraints on the speech content. A text-independent system generally requires more training and test data [2, 3] than a text-dependent one [4, 5] to maintain the same level of accuracy. Therefore, text-dependent system is preferable in many real-world applications where user's cooperation is possible.

The research on text-dependent SV has evolved a lot from traditional dynamic time warping based template matching method to deep learning recently. For benchmarking of technology process, standard speech corpora like RSR2015 and Red-Dots are designed [6,7]. It is found that the modeling techniques such as hierarchical multi-layer acoustic model (HiLAM) [6], unsupervised hidden Markov model (HMM)-universal background model (UBM) [8] i-vector/HMM [9] and j-vector [10] benefit from temporal information in speech. Further, deep learning techniques have greatly improved the ability of speaker characterization [11–16].

In human SV, we would like to have test samples as parallel to the unknown as possible to reduce the variability to a minimum due to language content. Text-dependent SV allows us to do just like that. Various studies show text-dependent SV is considered for performing two tasks, an SV as the main task, and an utterance verification as the subtask, where the two tasks are optimized separately or jointly in order to improve the SV objective. For example, the phonetic posteriorgrams derived using Gaussian mixture model (GMM) and deep neural network (DNN) frameworks are utilized to capture lexical information for text-dependent SV [17, 18]. One shows that DNN based speaker embedding benefits from lexical content information [19], others suggest that lexical information can be used in different ways to compensate the SV scores for performance gains [20–22].

Prior studies have underscored the importance of content modeling in text-dependent SV. While utterance verification has been well studied as part of speech recognition [23, 24], it has not been given sufficient attention in the context of SV. Some consider text-dependent SV as a combination of two independent systems, namely SV and utterance verification [25, 26]. In our previous work [27], text-dependent SV is formulated in a unified speaker-utterance verification (SUV) system as a multi-task learning implementation. This is inspired by human cognitive process where we interpret and decode speaker traits and linguistic content in a corroborative manner [28, 29]. For example, by paying special attention to particular sounds while knowing the linguistic content information, we verify the voice of a speaker; on the other hand, if we are familiar with a speaker, we tend to recognize his/her voice in a better way.

In the unified SUV system, we used a shared long short term memory (LSTM) network and two independent LSTM output layers, one for speaker identity and another for utterance identity [27]. While the previously proposed unified SUV framework is effective, the interaction between the two output layers is not explored. We believe that both speaker and utterance verification can benefit from each other by exploring the temporal interaction between them. This is also motivated by successful explorations in text-dependent SV that suggest the benefit of compensating lexical information [20–22]. Further, various attention models [30–32] project their possibility to focus on specific compensation or masking related to each task.

In this work, we propose an speaker-utterance dual attention, SUDA in short for performing both speaker and utterance verification. The attention mechanism for compensating irrelevant information for both tasks is derived by using masking operation. The masking for attention is estimated frame-by-frame, the attention mechanism establishes the temporal association between the speaker trait stream and the utterance content stream of LSTM output. In addition, we note that, as the attention mechanism is applied to both the branches (speaker and utterance) in the framework, it is referred to as dual attention. The studies in this work are conducted on RSR2015 corpus [6]. The contribution of this work lies in the use of a speaker-utterance dual attention in a single framework for performing both speaker and utterance verification.

The rest of the work is organized as follows. Section 2 describes the proposed speaker-utterance dual attention mechanism for speaker and utterance verification. The experiments are detailed in Section 3, followed by reporting of their results and analysis in Section 4. The paper is finally concluded in Section 5.
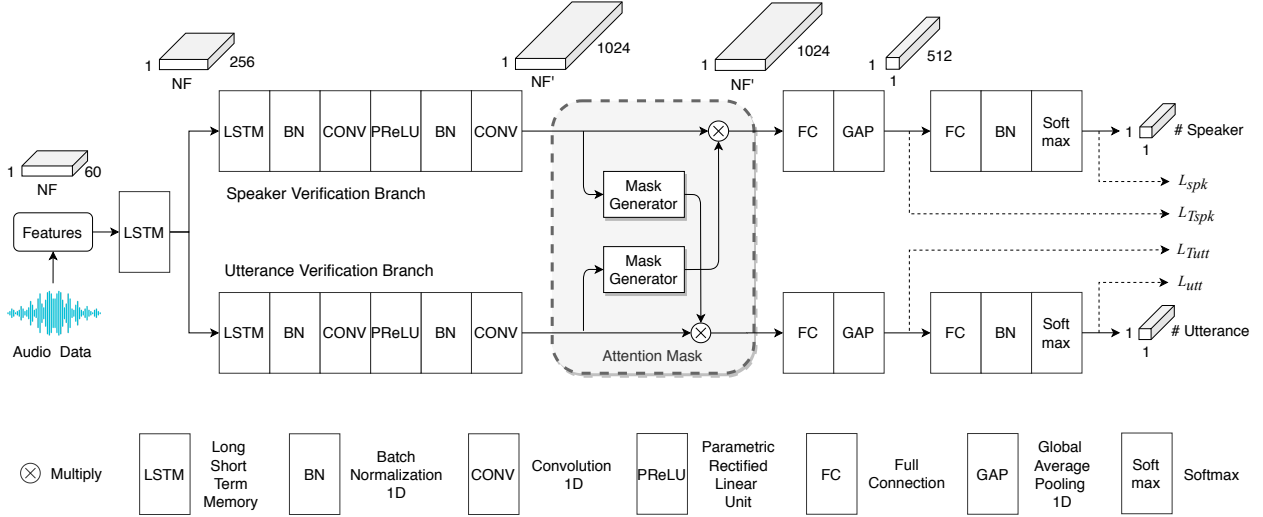
Figure 1: *Block diagram of proposed SUDA for speaker and utterance verification. It consists of a shared LSTM, two LSTM output layer, which interact each other through the attention mask network. The NF and NF' are number of frames before and after convolution.*

2. Speaker-Utterance Dual Attention

This section describes the proposed SUDA for speaker and utterance verification. As shown in Figure 1, the features extracted from raw audio data are fed into the LSTM based recurrent network to characterize the temporal dynamics. This system is an extension to our earlier work of unified SUV framework [27].

As presented in our earlier study [27], the first shared layer is common for speaker and utterance verification branches. Then, the hidden representation from the first shared layer is passed to the next two LSTM networks that focus on extracting valid information for each of the two sub-tasks, namely, speaker and utterance verification. We now discuss the improvisations introduced in this work using dual attention by masking.

2.1. Related Studies

Masking has been found to be effective to separate multiple sound sources from an audio mixture [33]. A more recent approach regards speech separation problem as a supervised learning that aims to discriminate different patterns such as, speech, speakers, and background noise, which are learned from training data [33]. Masking operation can be viewed as an attention mechanism, where the every single features from one branch are masked to attend the specific features from another branch. Attention based models have been applied successfully to various tasks. In computer vision research, there are three main strategies for attention: spatial attention, channel attention and mixed attention [34–36]. The spatial attention is used to emphasis the area that one is interested, while the channel attention is mainly for features recalibration in convolutional neural networks, such as the 'Squeeze-and-Excitation' block proposed in [35]. Again, the authors of [30] proposed an end-to-end framework with an attention model to combine the frame-level features, which acts as an alignment operation for SV studies. The attention model designed as a part of the speaker embedding network, is used to calculate the weighted mean of the frame-level feature maps to derive speaker embedding with a better discriminative speaker characteristics [31, 32].

2.2. Attention Mask

In the proposed SUDA, the convolution layers are used to extract the feature maps from the hidden representation obtained after LSTM as shown in Figure 1. It has the objective to map the hidden representation to higher dimensional space to compensate the mutual information by attention masks. The dynamic masks can learn the information from feature maps obtained from the convolutional layers. We use sigmoid non-linearity to limit the activation from the mask. The dynamic masking is performed by learning the parameters from feature maps and the activating sigmoid function [37]. The masking is formulated as:

$$\text{mask}_s = 1 - \text{Sigmoid}(fm_u) \tag{1}$$

$$\text{mask}_u = 1 - \text{Sigmoid}(fm_s) \tag{2}$$

where the fm_u and fm_s represent the feature maps from utterance verification and SV branch, respectively, while the mask_u and mask_s indicate the dynamic masks for the corresponding branch. These masks are then multiplied by the corresponding stream of feature maps to suppress irrelevant information in the respective stream of feature maps. The parameters in the masks are not fixed during the training or inference phase. They are derived according to the input audio data.

The operations performed so far produce the framewise representations. In order to pool the information across the utterance, global average pooling (GAP) is performed. Next, the fully connected (FC) layers are used to perform both speaker and utterance verification tasks.

In this work, the attention mask is applied on the every feature map through sigmoid operation than the conventional softmax operation. Further, the weights of our attention masks are obtained from the branch of another task. We note that the weights of the attention mask are tuned for different utterances and speaker-specific information.

3. Experiments

We now discuss the database and experimental setup for the studies in the following subsections.

Table 1: *Performance in EER (%) for proposed SUDA in comparison to existing systems on RSR2015 Part I corpus.*

System	Male						Female					
	Development			Evaluation			Development			Evaluation		
	TW	IC	IW	TW	IC	IW	TW	IC	IW	TW	IC	IW
i-vector [6]	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 [6]	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-spk-utt [27]	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 [27]	0.470	1.590	0.101	0.293	1.757	0.039	1.176	4.323	0.178	0.375	2.009	0.068
Utt-comp [22]	-	1.460	-	-	0.960	-	-	1.640	-	-	0.730	-
Utt-comp-uf [22]	-	1.460	-	-	0.960	-	-	1.460	-	-	**0.720**	-
mod-SUV	**0.202**	0.952	0.034	0.107	1.093	0.020	0.475	2.055	0.071	0.182	1.373	0.034
Proposed: SUDA	**0.202**	**0.728**	**0.022**	**0.068**	**0.722**	**0.010**	**0.297**	**1.449**	**0.024**	**0.125**	0.863	**0.023**

3.1. Database

The RSR2015 corpus is used for the studies in this work [6]. It contains 300 speakers data from 143 female and 157 male speakers. Further, the corpus is divided into three different parts based on the nature of the fixed phrases. The Part I includes 30 fixed phrase utterances of 3-4 seconds duration, whereas the Part II has 30 fixed short commands of 1-2 seconds duration. Similarly, the Part III contains the random digit based five or ten digit sequences. Again, there are 9 sessions for each phrase from all the speakers. Out of those, the first, fourth and seventh sessions are used for speaker enrollment and the remaining for testing as per the RSR2015 evaluation protocol [6].

The RSR2015 corpus has three subsets, which are background, development and evaluation set to evaluate the system performance [6]. The test trials are grouped under four categories based on the test speaker and phrase labels, which are *Target Correct* (TC), *Impostor Correct* (IC), *Target Wrong* (TW) and *Impostor Wrong* (IW). They further constitute three test conditions, where each of those conditions consider TC as target trials and the remaining three categories as non-target trials, respectively. The performance is reported in terms of equal error rate (EER). In this work, we consider Part I and Part II of RSR2015 as they are suitable for both speaker and utterance verification.

3.2. Experimental Setup

The speech utterances are processed 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 extracted features are normalized by cepstral mean and variance normalization using utterance level mean and variance statistics.

In addition, we apply the feature level triplet loss to further increase the intra-class distance and reduce inter-class distance. It is applied on the 512-dimensional feature vector at the step after 1D global average pooling as shown in Figure 1. We calculate triplet loss in both the branches, i.e., SV and utterance verification. In addition, the negative and positive samples in the training batch for each iteration is searched. The batch size of each iteration is set to be 128 for all the experiments and these samples are chosen randomly. The loss of proposed framework is calculated by following formula:

$$L_{total} = L_{Tspk} + L_{Tutt} + L_{spk} + L_{utt} \qquad (3)$$

---
[1] http://kaldi-asr.org/
[2] https://pytorch.org/

where the L_{Tspk} and L_{Tutt} are triplet loss, while L_{spk} and L_{utt} are negative log likelihood loss. The subscripts spk and utt represent the SV and utterance verification branches in SUDA, respectively.

The learning rate, optimizer, LSTM's hidden layer and scoring follow the same configurations as that in our previous work of unified SUV [27]. We adopt PyTorch[2] toolkit for the implementation. We fixed seeds empirically to 2020 in our studies. In contrast to the previous unified SUV, we add 1D convolution layer to the network, where kernel size of convolution layer is 5, while the padding and the stride are 0 and 1, respectively. The activation function between convolutional layers is parametric rectified linear unit (PReLU) [38]. Further, we note that in order to observe the impact of attention mask, we also conduct the experiments without attention based on masking operation block. We refer to the system with this setup as mod-SUV for comparing to previous unified SUV [27] in this work.

4. Results and Discussions

We consider HiLAM and i-vectors as two basic common reference systems for the studies [6]. Further, as the work advocates on compensation network, we consider joint speaker utterance (joint-spk-utt) [41] and utterance compensation (utt-comp) [22] frameworks for comparison. We note that joint speaker utterance models speaker and utterance information jointly [41], whereas the utterance compensation framework compensates the utterance information after jointly modeling speaker and utterance characteristics [22]. Further, the utterance compensation framework has another variant with utterance factor (utt-comp-uf). We note that all these works used for comparison targets for only SV studies. The unified SUV proposed in our previous work for performing both speaker and utterance verification is also used as reference system [27].

Table 1 shows the performance comparison of our proposed SUDA to the systems discussed above on Part I of RSR2015 corpus. The performance of systems compared are quoted from the previously published results. We find that HiLAM system performs better than the i-vectors systems due to use of temporal knowledge. Further, the joint speaker utterance model outperforms HiLAM system for IC test trial condition, while it performs poorly in TW and IW trial conditions related to utterance verification. Compared to these two systems, our previous work, unified SUV takes advantage of LSTM to capture temporal dynamics, greatly reduces the error rate of TW and IW trial conditions. In the utterance compensation system, we observe compensating utterance information leads to a good performance in IC trial condition that further improves in addition

Table 2: *Performance in EER (%) for proposed SUDA in comparison to existing systems on RSR2015 Part II corpus.*

System	Male						Female					
	Development			Evaluation			Development			Evaluation		
	TW	IC	IW	TW	IC	IW	TW	IC	IW	TW	IC	IW
i-vector [6]	5.410	13.750	2.500	4.390	11.260	1.810	6.940	12.730	2.860	5.160	15.270	3.050
HiLAM [6]	6.140	10.580	3.030	4.420	8.380	1.710	4.620	6.660	1.290	3.710	7.950	1.450
Joint-spk-utt	10.804	4.096	2.715	9.929	4.190	2.286	10.220	3.482	2.179	7.797	3.382	1.816
Utt-comp [22]	-	4.160	-	-	3.610	-	-	4.030	-	-	2.850	-
Utt-comp-uf [22]	-	4.160	-	-	3.610	-	-	3.860	-	-	2.790	-
mod-SUV	1.394	3.757	0.279	1.015	3.591	0.176	1.833	4.862	**0.272**	0.851	3.563	0.102
Proposed: SUDA	**1.382**	**2.698**	**0.245**	**0.878**	**2.400**	**0.127**	**1.360**	**3.359**	0.296	**0.624**	**2.020**	**0.057**

Table 3: *Performance in EER (%) of different systems on the evaluation set of RSR2015 Part I. Here, j-vector: j-vector with cosine similarity, Joint Bayesian: 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, J3: joint training of j-vector extractor and joint Bayesian, and use the Siamese network output for verification, RACNN-LSTM: raw audio convolutional neural network with LSTM, i-vector + s-vector: the system concatenating i-vector and s-vector directly, i-s-vector: the system concatenating the last-step hidden output of s-vector and corresponding i-vector (s-vector extracted either from LSTM or Bidirectional LSTM (BLSTM)).*

System	TW	IC	IW
j-vector [12]	3.14	7.86	0.95
Joint Bayesian [15]	0.03	3.61	0.02
J2 [15]	**0.02**	2.81	0.02
J3 [15]	**0.02**	2.42	0.02
RACNN-LSTM [39]	-	3.63	-
Unified SUV [27]	0.46	2.41	0.06
i-vector + s-vector [40]	0.28	1.13	0.03
i-s-vector (LSTM) [40]	0.17	1.98	0.03
i-s-vector (BLSTM) [40]	0.11	1.72	0.02
mod-SUV	0.15	1.14	0.02
Proposed: SUDA	0.13	**0.62**	**0.01**

Table 4: *Performance in EER (%) for SV and utterance verification (UV) on Part I of RSR2015 evaluation set.*

System	Male		Female	
	SV	UV	SV	UV
Unified SUV [27]	1.796	0.021	1.918	**0.011**
mod-SUV	1.132	0.010	1.362	**0.011**
Proposed: SUDA	**0.741**	**0.005**	**0.840**	**0.011**

of the utterance factor [22]. However, it did not explore on compensating speaker information for utterance verification unlike this paper, hence, the results of TW and IW trials are not investigated in [22].

In this work, as mentioned in Section 3.2, mod-SUV is a modified framework of unified SUV. We can observe from Table 1 that mod-SUV has significantly improved SV performance in all three test trial conditions over the existing unified SUV framework. Further, our proposed SUDA, which focuses on the required speaker and utterance information by imposing dual attention, outperforms most of the other systems for all three test trials conditions except, IC test trial condition of evaluation female set.

Table 2 reports the performance of various systems on Part II of RSR2015 database. The performance trend of various systems remains similar to that observed in case of Part I of RSR2015 database. The proposed SUDA again outperforms all other systems, showing effectiveness of attention based dual compensation in various trial conditions.

We now compare our proposed SUDA framework with other deep learning systems. We combine both male and fe-

male data of RSR2015 Part I to match with the evaluation protocol followed in [12, 15, 39, 40] for comparison with other research studies. We observe from Table 3 that the mod-SUV has a comparable performance to other systems in the IC trial condition. Further, with the use of attention masks, the proposed SUDA achieves a significant improvement on the IC trial condition, and at the same time improves performance for IW and TW trial conditions.

Finally, as discussed in our earlier work [27], we can adjust and tune the scores from speaker and utterance verification branch to show a security trade-off during scoring. The studies related to this reported in Table 4 show that the proposed SUDA outperforms previous unified SUV and the current mod-SUV, that highlights the gain provided by dual attention mechanism.

5. Conclusions

This work proposes a novel speaker-utterance dual attention (SUDA) mechanism for speaker and utterance verification. We used LSTM based models with two branches in a unified framework that consider attention masks for temporal interaction between speaker trait stream and utterance content stream that helps to suppress the irrelevant information for both tasks. The studies conducted on RSR2015 corpus reveal the importance of the proposed SUDA in comparison to existing approaches to work effectively for both speaker and utterance verification simultaneously. The framework also leverages a user for using it according to the security need of the intended application. The future work will focus on extending attention masking for prompted digit based SV.

6. Acknowledgements

7. References

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

[2] R. K. Das, *Speaker verification using sufficient train and limited test data*. PhD thesis, Septemebr 2017.

[3] R. K. Das, S. Jelil, and S. R. M. Prasanna, "Significance of constraining text in limited data text-independent speaker verification," in *SPCOM*, Bangalore, India, 2016, pp. 1–5.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

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

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

[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] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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*, Les Sables d'Olonne, France, 2018, pp. 365–371.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

[23] 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.

[24] 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.

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

[26] H. Zeinali, L. Burget, H. Sameti, and H. Cernocky, "Spoken passphrase verification in the i-vector space," in *Odyssey*, Les Sables d'Olonne, France, 2018, pp. 372–377.

[27] T. Liu, M. Madhavi, R. K. Das, and H. Li, "A unified framework for speaker and utterance verification," in *Interspeech*, Graz, Austria, 2019, pp. 4320–4324.

[28] 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.

[29] 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.

[30] S.-X. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-end attention based text-dependent speaker verification," in *SLT Workshop*, San Juan, Puerto Rico, 2016, pp. 171–178.

[31] F. R. rahman Chowdhury, Q. Wang, I. L. Moreno, and L. Wan, "Attention-based models for text-dependent speaker verification," in *ICASSP*, Calgary, Alberta, Canada, 2018, pp. 5359–5363.

[32] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification." in *Interspeech*, Hyderabad, India, 2018, pp. 3573–3577.

[33] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[34] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *NIPS*, Montreal, Canada, 2015, pp. 2017–2025.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, Salt Lake City, Utah, USA, 2018, pp. 7132–7141.

[36] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *CVPR*, Honolulu, Hawaii, USA, 2017, pp. 3156–3164.

[37] C. Shan, J. Zhang, Y. Wang, and L. Xie, "Attention-based end-to-end speech recognition on voice search," in *ICASSP*, Calgary, Alberta, Canada, 2018, pp. 4764–4768.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[39] J.-W. Jung, H.-S. Heo, I.-H. Yang, H.-J. Shim, and H.-J. Yu, "A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result," in *ICASSP*, Calgary, Alberta, Canada, 2018, pp. 5349–5353.

[40] S. Wang, Y. Qian, and K. Yu, "What does the speaker embedding encode?" in *Interspeech*, Stockholm, Sweden, 2017, pp. 1497–1501.

[41] 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, USA, 2016, pp. 415–419.