Neural Acoustic-Phonetic Approach for Speaker Verification With Phonetic Attention Mask

Tianchi Liu, *Graduate Student Member, IEEE*, Rohan Kumar Das, *Senior Member, IEEE*,
Kong Aik Lee, *Senior Member, IEEE*, and Haizhou Li, *Fellow, IEEE*

***Abstract*—Traditional acoustic-phonetic approach makes use of both spectral and phonetic information when comparing the voice of speakers. While phonetic units are not equally informative, the phonetic context of speech plays an important role in speaker verification (SV). In this paper, we propose a neural acoustic-phonetic approach that learns to dynamically assign differentiated weights to spectral features for SV. Such differentiated weights form a phonetic attention mask (PAM). The neural acoustic-phonetic framework consists of two training pipelines, one for SV and another for speech recognition. Through the PAM, we leverage the phonetic information for SV. We evaluate the proposed neural acoustic-phonetic framework on the RSR2015 database Part III corpus, that consists of random digit strings. We show that the proposed framework with PAM consistently outperforms baseline with an equal error rate reduction of 13.45% and 10.20% for female and male data, respectively.**

***Index Terms*—Speaker verification, text-dependent, attention, masking, phonetic information, prompted digit recognition.**

I. Introduction

SPEAKER verification (SV) seeks to verify the claimed identity of a speaker [1]. It can be generally categorized into text-dependent and text-independent tasks [2], [3]. The former typically requires less training and test data than the latter to maintain the same level of performance [4], [5]. Text-dependent SV task allows us to compare utterances of the same phonetic context [6], [7], or random word sequences coming from a fixed vocabulary [8], [9]. With random sequences, such as random digit strings, an SV system is less vulnerable to replay attacks [8], [10], [11]. In this work, we study a neural acoustic-phonetic approach for SV of random digit strings in RSR2015 Part III database [12].

Speech signals contain both speaker traits and phonetic context [13], [14], of which phonetic context is the carrier of speaker traits. Humans identify voice print more accurately when they associate phonetic context with acoustic information [15]. Studies show that humans take advantage of language-specific knowledge of speech phonology to facilitate speaker recognition [16]. Similarly, SV benefits from phonetically-aware speaker traits [14], [17]–[26]. For example, a c-vector architecture is studied to introduce frame-level phonetic information into segment-level speaker embedding [14]. In [17], phonetic features are projected into multi-channel feature maps. Along similar direction, PacNet [18] uses coupled stem to jointly learn acoustic features and transform frame-level ASR bottleneck feature. In [25], a speaker-utterance dual attention (SUDA) is applied to learn the interaction between speaker and utterance information streams in a unified framework for text-dependent SV.

Along this line of thought, we believe that the speaker traits carried by different phonetic contexts are not equally informative. The question is how to model the interaction between speaker and phonetic features. In most of the prior studies, speaker and phonetic features are often encoded independently, resulting in different feature spaces. Therefore, the acoustic-phonetic interaction is not localized at the time-frequency bins of the speaker feature map.

In this work, we propose a neural acoustic-phonetic approach with phonetic attention mask (PAM) that learns to dynamically assign differentiated weights to speaker feature map. The PAM is different from SUDA [25] implementation, where acoustic and phonetic feature maps are fused with a constant weight. In addition, we propose a network architecture that is optimized for phonetically informed SV.

II. Neural Acoustic-Phonetic Approach

We propose a dual-pathway network, as illustrated in Fig. 1, that facilitates the joint optimization between a speaker classification and a speech recognition task, therefore, the interaction between the two.

Manuscript received October 12, 2021; revised December 26, 2021; accepted January 3, 2022. Date of publication January 13, 2022; date of current version March 24, 2022. This work was supported in part by Programmatic under Grant A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain); in part by the Human-Robot Interaction Phase 1 under Grant 192 25 00054; in part by the National Research Foundation, Prime Minister's Office, Singapore, under the National Robotics Programme; and in part by the Agency for Science and Engineering Research Council (A*STAR) under its Feasibility Study Scheme under Grant FS-2021-001, in part by A STAR under its RIE2020 Advanced Manufacturing and Engineering Domain (AME) under Grant A1687b0033. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Rosangela Coelho. *(Corresponding authors: Liu Tianchi; Rohan Kumar Das.)*

Tianchi Liu is with the Institute for Infocomm Research, A*STAR, Singapore 138632, and also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: e0320839@u.nus.edu).

Rohan Kumar Das is with Fortemedia, Singapore 138589 (e-mail: ecerohan@gmail.com).

Kong Aik Lee is with the Institute for Infocomm Research, A*STAR, Singapore 138632 (e-mail: kongaik.lee@gmail.com).

Haizhou Li is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: haizhou.li@u.nus.edu).

Digital Object Identifier 10.1109/LSP.2022.3143036

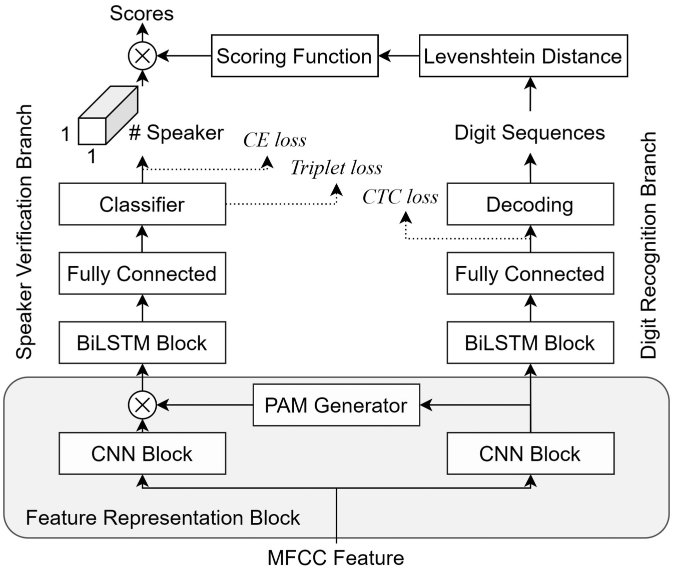


Fig. 1. Block diagram of the proposed neural acoustic-phonetic approach with phonetic attention mask for SV. $\otimes$ indicates element-wise multiplication.

A. Phonetic Attention Mask (PAM)

We propose a neural acoustic-phonetic approach with a phonetic attention mask (PAM) that learns to dynamically assign such weights during run-time inference. The neural solution consists of two pathways that learn to encode speaker and linguistic content separately, and a PAM that interacts between the two pathways.

The digit recognition pathway in Fig. 1 can be specifically optimized for providing the representations of phonetic information. A PAM generator bridges the two pathways by producing a mask of the same size as the speaker feature map, as shown in Fig. 1, according to the phonetic context characterized by the pre-trained digit recognizer. We regard the dual-pathway network without PAM generator as a reference baseline for the proposed neural acoustic-phonetic approach.

We train the SV pathway and the PAM generator jointly. The PAM generator includes one layer of 1D CNN with kernel size of 1 and a sigmoid non-linearity layer. The former is used to re-weight each element on the feature map according to phonetic information, while the latter can restrict the activation. The PAM generator generates an attention map which assigns a differentiated weight dynamically to each time-frequency bin of the speaker feature map as follows,

$$PAM = 1 - Sigmoid(Conv1D(F_{dgt})) \tag{1}$$

where $Conv1D$ denotes a 1D CNN. F_{dgt} is the speech feature map from the digit recognition CNN block.

The CNN blocks of the two pathways are specifically designed to have the same architecture so that the speech feature map F_{dgt} and speaker feature map F_{spk} are of the same dimension. During run-time inference, a PAM is generated dynamically according to input speech context, which modulates the speaker feature map F_{spk} as follows,

$$F'_{spk} = F_{spk} \otimes PAM \tag{2}$$

where $\otimes$ denotes element-wise multiplication.

B. Acoustic-Phonetic Feature Concatenation (concat)

We further study an alternative approach to acoustic-phonetic feature representation. The idea of [17] is intuitive and effective. To fairly compare it with the proposed PAM, we reimplement its basic idea based on our baseline system as a contrastive system by concatenating phonetic features with acoustic features to fuse the two features, that is referred to as '*concat*' approach from here on.

As shown in Fig. 2(b), the speech feature map F_{dgt} is first processed by a 1D CNN with kernel size of 1 to reduce the number of channels to N, then reshaped into the same dimension as the MFCC features by two layers of fully connected layer. In this way, the reshaped features are considered to encode the phonetic information, thus also referred to as phonetic features. Finally, the phonetic and acoustic features are concatenated to form $N + 1$ channels as the input to SV.

C. Training Strategy

The digit recognition model is first trained with connectionist temporal classification (CTC) loss [27] for digit recognition. In the *baseline* model, the digit recognition and the SV pathways are trained independently. In PAM or *concat* model, the weights of the well-trained digit recognition model are fixed during the training of SV branch. In PAM, the PAM generator and the SV model are bridged as in Fig. 2(a) for joint training; while in *concat*, the SV model is trained on acoustic-phonetic concatenated features as in Fig. 2(b).

Cross entropy (CE) loss and embedding level triplet loss are adopted for training the SV model. Specifically, the triplet loss is applied on the 512-dimensional embedding vector after 1D global average pooling in the classifier. The positive and negative samples for calculating triplet loss are randomly sampled in each batch training iteration.

III. Experiments

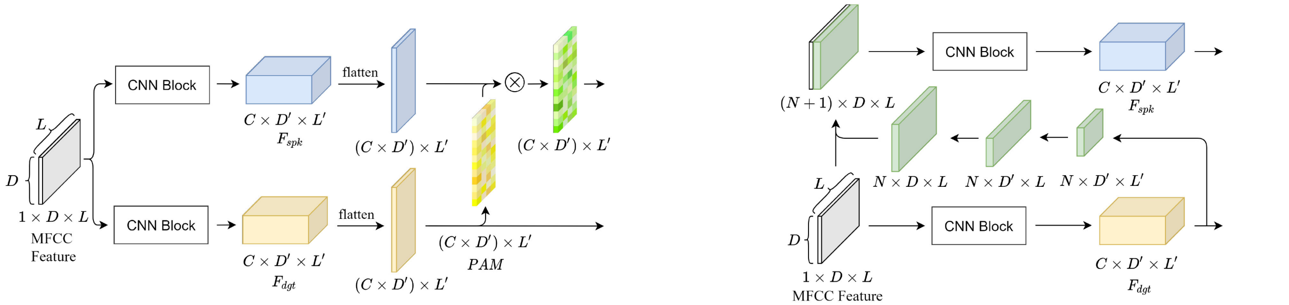
A. Database

We perform experiments on RSR2015 Part III [12], which has a close set of speakers. For the SV model, we use three strings each consisting of all 10 digits in a random order as training set, and quasi-random 5-digit strings as the test set following the protocol of RSR2015 data description. In addition, the background set of RSR2015 Part III, and the 5-digit SV test set are used to train a digit recognition model.

The 5-digit SV test trials can be grouped into four categories by pairing speaker identity and speech content, namely, Target Correct (TC), Impostor Correct (IC), Target Wrong (TW), and Impostor Wrong (IW), as described in [12]. Among the four types of trials, TC is seen as target trials, while the rest are used as non-target, of which each constitutes a test condition.

B. Experimental Setup

The speech data of RSR2015 database are processed with 20 ms frame size and 10 ms shift to extract 60-dimensional (20-base + 20-Δ + 20-$\Delta\Delta$) MFCC features. The MFCC features are further normalized with cepstral mean and variance



(a)

(b)

Fig. 2. The feature representation block of the dual-pathway framework. (a) The proposed neural acoustic-phonetic approach with a phonetic attention mask. (b) The *concat* approach that fuses phonetic features and acoustic features as the input to the speaker verification pathway. (a) Details the gray panel in Fig. 1. The input mel frequency cepstral coefficients (MFCC) is represented as a $1 \times L \times D$ feature map with 1 channel, L frames and D dimensional feature frame. C, L' and D' are number of channels, frames and feature dimensions after convolution, respectively. $\otimes$ Indicates element-wise multiplication.

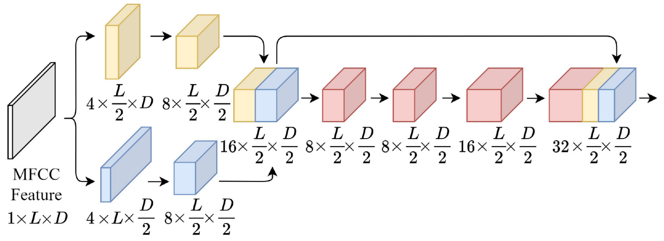

$4\times\frac{L}{2}\times D$
$8\times\frac{L}{2}\times\frac{D}{2}$
$16\times\frac{L}{2}\times\frac{D}{2}$
$8\times\frac{L}{2}\times\frac{D}{2}$
$8\times\frac{L}{2}\times\frac{D}{2}$
$16\times\frac{L}{2}\times\frac{D}{2}$
$32\times\frac{L}{2}\times\frac{D}{2}$
MFCC
Feature
$1\times L\times D$
$4\times L\times\frac{D}{2}$
$8\times\frac{L}{2}\times\frac{D}{2}$

Fig. 3. The network architecture of the CNN block in the experiments. The MFCC is represented as a $1 \times L \times D$ feature map with 1 channel, L frames and D dimensional feature frame. The 3D-block represents the shape of features.

normalization (CMVN) by utterance level mean and variance statistics [30].

In all systems, as shown in Fig. 1, we inherit the design of Convolutional, and Long Short-Term Memory Deep Neural Network (CLDNN) pipeline [31], which is widely adopted [32]–[37]. The detailed architecture of CNN block for feature extraction used in our experiments is shown in Fig. 3. The activation function of parametric rectified linear unit (PReLU) [38], batch normalization [39] and dropout [40] are employed in between convolutional layers. The classifier module in the SV branch contains 1D global average pooling, fully connected layer and batch normalization operation. In the digit recognition branch, a beam search algorithm is implemented to decode the digit sequence [41].

C. System Training

All systems are trained on the same dataset for fair comparison. The batch size is fixed to 128 for all experiments. The random seeds for the initialization of weights are empirically fixed to 50, 100, 500 and 1,000 in this work, and each experiment is repeated four times to report the mean. In addition, the learning rate is set to 0.0003, and the dimension of hidden layer of BiLSTM is 512. We set the number of channels, N, to 2. Therefore, we form the acoustic-phonetic features of $N + 1 = 3$ channels as the input to the SV task.

D. System Description

- Acoustic-only baseline: The acoustic-only baseline only includes the SV branch of the baseline system.

TABLE I
PERFORMANCE COMPARISON IN EER (%) BETWEEN THE PROPOSED NEURAL ACOUSTIC-PHONETIC APPROACH AND OTHER STATE-OF-THE-ART SYSTEMS FOR TC-IC TRIAL OF RSR2015 PART III

System	Feature	Female/Male
x-vector [8]	MFCC	2.88% / 2.71%
DNN/GMM-MAP [10]	FBank	2.68% / 2.08%
Bottleneck DNN i-vector [40]	MFCC	2.54% / 1.98%
Bottleneck DNN i-vector [40]	Tandem	1.84% / 1.81%
S-U-J + s-norm [9]	MFCC	1.84% / 1.57%
Double joint Bayesian + s-norm [41]	MFCC	1.66% / 1.48%
Uncertainty norm [8]	MFCC	1.77% / 1.52%
Uncertainty norm [8]	MFCC + PLP	**1.65% / 1.45%**
Acoustic-only baseline	MFCC	1.71% / 1.47%
Acoustic-phonetic system with PAM	MFCC	**1.48% / 1.32%**

- Acoustic-phonetic baseline: this system has the same architecture as in Fig. 1, except that there is no PAM generator to interact between the pathways.
- Acoustic-phonetic system with *PAM:* The proposed neural acoustic-phonetic approach with *PAM* as shown in Fig. 1.
- Acoustic-phonetic system with *Concat:* The contrastive system which concatenates phonetic features with acoustic features for early fusion, as described in Section II-B, in place of *PAM*.

E. Acoustic-Phonetic Scoring

We derive a total score from both acoustic and phonetic pathways. The Levenshtein distance [42] is computed between a predicted digit sequence and its reference, which is further transformed by a scoring function as shown in Fig. 1, and formulated next.

$$score_{dgt} = Sigmoid(\gamma - 2 * lev_{dgt}) \tag{3}$$

where lev_{dgt} denotes the Levenshtein distance, while $\gamma = 5$ is the number of digits in the reference digit strings. The $score_{spk}$ is obtained from the output of the speaker verification pathway, which represents the posterior probability of each speaker. We adopt weighting factor α [43] to fuse $score_{spk}$ and $score_{dgt}$, and derive a total score $score_{total}$ as (4).

$$score_{total} = \alpha \log(score_{spk}) + (1 - \alpha) \log(score_{dgt}) \tag{4}$$

where we set $\alpha = 0.7$ empirically in the experiments.

TABLE II
SV PERFORMANCE ON RSR2015 PART III CORPUS IN EER (%), AND TOP-1 ACCURACY (%), WE REPORT BOTH OVERALL EER AND BY TRIAL CATEGORY, I.E. TC-IC AND TC-IW

System	Female				Male			
	TC-IC	TC-IW	Overall ↓	Top-1 Acc ↑	TC-IC	TC-IW	Overall ↓	Top-1 Acc ↑
Acoustic-only baseline	1.71%	1.73%	1.73%	92.13%	1.47%	1.50%	1.50%	93.24%
Acoustic-phonetic system with Concat	1.65%	1.68%	1.68%	92.48%	1.39%	1.42%	1.42%	93.77%
Acoustic-phonetic system with PAM	**1.48%**	**1.50%**	**1.50%**	**93.26%**	**1.32%**	**1.33%**	**1.33%**	**93.90%**
(Upper Bound)	(1.42%)	(1.45%)	(1.45%)	(93.22%)	(1.27%)	(1.29%)	(1.29%)	(94.33%)

IV. RESULTS AND DISCUSSIONS

A. Speaker Verification

In Table I, we compare the performance in terms of equal error rate (EER) between the proposed acoustic-phonetic approach and other state-of-the-art systems on RSR2015 Part III for the TC-IC trials, where the speech content is always correct. The EER is calculated only by $score_{spk}$ to purely evaluate the SV performance. We observe that the acoustic-only baseline achieves a comparable result with other competing systems, and the acoustic-phonetic system with PAM achieves a performance gain of EER reduction by 13.45% and 10.20% for female and male data, respectively. The performance gain is attributed due to the proposed PAM mechanism.

We now further perform ablation study on the proposed acoustic-phonetic system with PAM, as summarized in Table II, in terms of EER and Top-1 Accuracy. As we evaluate the EER by $score_{spk}$, we only report two trial categories, namely TC-IC and TC-IW, where TC are the target set, and IC and IW are two different non-target sets. We observe that the proposed neural acoustic-phonetic approach significantly improves the performance in all cases over the acoustic-only approach.

Now we would like to compare the proposed PAM with a contrast system $concat$ which also benefits SV from phonetically-aware speaker traits. The $concat$ system concatenates the phonetic information based transformed features with MFCC to consider as input to the baseline. We find from Table II that our baseline framework can perform better with PAM than $concat$ approach for all of the evaluation indicators by introducing phonetic information into SV more effectively. This may be due to the fact that proposed attention mask can utilize the phonetic information to assist SV task by re-weighting each time-frequency bin effectively.

Further, we would like to examine the effect of speech recognition accuracy. To this end, we add the 5-digit based test data into the training set for digit recognition model training. In this way, it is expected that the digit recognition model fits well with the test set to obtain high quality phonetic information feature maps, that gives us the upper bound performance. We can observe from Table II that the acoustic-phonetic system with PAM approximates the performance of the upper bound performance.

TABLE III
MULTI-TASK PERFORMANCE ON RSR2015 PART III CORPUS IN EER (%)

System	Female			Male		
	TC-TW	TC-IC	TC-IW	TC-TW	TC-IC	TC-IW
Acoustic-Phonetic	1.18%	1.74%	**0.08%**	1.41%	1.62%	0.15%
+ Concat	1.21%	1.65%	**0.08%**	1.45%	1.53%	**0.12%**
+ PAM	**1.13%**	**1.51%**	0.09%	**1.36%**	**1.46%**	0.18%
(Upper Bound)	(1.06%)	(1.42%)	(0.08%)	(1.19%)	(1.27%)	(0.07%)

TABLE IV
COMPARISON OF SV PERFORMANCE OF PROPOSED ACOUSTIC-PHONETIC SYSTEM WITH PAM BETWEEN CASES THAT DIGITS ARE CORRECTLY OR WRONGLY PREDICTED

Digit Recognition Prediction	EER of SV	
	Male	Female
Correct	1.17	1.41
Wrong	2.98	1.85

B. Multi-Task Experiments

In Table III, we compare four systems that are trained with both SV and digit recognition tasks. Here, the EER results for each trial category, namely TC-TW, TC-IC, and TC-IW, are evaluated with $score_{total}$ in (4) for both tasks at the same time. The performance trend of various systems mostly remains similar to those evaluated with $score_{spk}$ in Table II.

It is worth noting that, unlike those in TC-IC trials, the digits strings in TC-TW and TC-IW categories are having incorrect speech content. Comparing the results in Table II and Table III, we are encouraged to see that, by evaluating both speech content and speaker scores, we achieve better results for TC-IC, TC-TW and TC-IW trials.

C. Digit Recognition

Finally, we would like to investigate how digit recognition performance impacts the SV results. We divided the samples into two groups based on the results of digit recognition and named them 'correct' and 'wrong' in Table IV. We only consider the case that when all predicted 5 digits are consistent with the prompted digits as 'correct'. From Table IV, we observe that when digits are correctly recognized by the digit recognition model, the performance of SV obviously outperforms. This may benefit from the higher quality of phonetic information provided by the digit recognition model in the case of correct digits recognition.

V. CONCLUSION

We propose a novel neural acoustic-phonetic approach for text-dependent SV with random digit strings, where we adopt a phonetic attention mask (PAM) to model the interaction between the acoustic and phonetic pathways. Studies on RSR2015 Part III reveal that we greatly benefit from the proposed PAM. In addition, the upper bound study shows that PAM is capable of utilizing the phonetic information very effectively.

REFERENCES

[1] J. P. Campbell, “Speaker recognition: A tutorial,” *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[2] M. Hèbert, “Text-dependent speaker recognition,” in *Springer Handbook of Speech Processing*. Berlin, Germany: Springer, 2008, pp. 743–762.

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

[4] R. K. Das and S. R. M. Prasanna, “Speaker verification from short utterance perspective: A review,” *IETE Tech. Rev.*, vol. 35, no. 6, pp. 599–617, 2018.

[5] R. K. Das, “Speaker verification using sufficient train and limited test data,” Ph.D dissertation, Indian Institute of Technology Guwahati, Sep. 2017.

[6] 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 *Proc. Interspeech*, 2016, pp. 415–419.

[7] R. K. Das, M. Madhavi, and H. Li, “Compensating utterance information in fixed phrase speaker verification,” in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2018, pp. 1708–1712.

[8] N. Maghsoodi, H. Sameti, H. Zeinali, and T. Stafylakis, “Speaker recognition with random digit strings using uncertainty normalized HMM-based i-vectors,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 11, pp. 1815–1825, Nov. 2019.

[9] Z. Shi, H. Lin, L. Liu, and R. Liu, “Latent factor analysis of deep bottleneck features for speaker verification with random digit strings,” in *Proc. Interspeech*, 2018, pp. 1081–1085.

[10] Y. Liu, L. He, W.-Q. Zhang, J. Liu, and M. T. Johnson, “Investigation of frame alignments for GMM-based digit-prompted speaker verification,” in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2018, pp. 1467–1472.

[11] S. Novoselov, O. Kudashev, V. Shchemelinin, I. Kremnev, and G. Lavrentyeva, “Deep CNN based feature extractor for text-prompted speaker recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5334–5338.

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

[13] R. K. Das, S. Jelil, and S. R. M. Prasanna, “Significance of constraining text in limited data text-independent speaker verification,” in *Proc. Int. Conf. Signal Process. Commun.*, 2016, pp. 1–5.

[14] Y. Liu, L. He, J. Liu, and M. T. Johnson, “Introducing phonetic information to speaker embedding for speaker verification,” *EURASIP J. Audio, Speech, Music Process.*, vol. 2019, no. 1, pp. 1–17, 2019.

[15] J. M. Zarate, X. Tian, K. J. Woods, and D. Poeppel, “Multiple levels of linguistic and paralinguistic features contribute to voice recognition,” *Sci. Rep.*, vol. 5, no. 1, pp. 1–9, 2015.

[16] T. K. Perrachione, S. N. D. Tufo, and J. D. Gabrieli, “Human voice recognition depends on language ability,” *Science*, vol. 333, no. 6042, pp. 595–595, 2011.

[17] T. Zhou, Y. Zhao, J. Li, Y. Gong, and J. Wu, “CNN with phonetic attention for text-independent speaker verification,” in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2019, pp. 718–725.

[18] S. Zheng, Y. Lei, and H. Suo, “Phonetically-aware coupled network for short duration text-independent speaker verification,” in *Proc. Interspeech*, 2020, pp. 926–930.

[19] A. S. Park, “ASR dependent techniques for speaker recognition,” Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2002.

[20] B. Baker, R. Vogt, and S. Sridharan, “Gaussian mixture modelling of broad phonetic and syllabic events for text-independent speaker verification,” in *Proc. Interspeech*, 2005, pp. 2429–2432.

[21] Y. Tian, L. He, M. Cai, W.-Q. Zhang, and J. Liu, “Deep neural networks based speaker modeling at different levels of phonetic granularity,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5440–5444.

[22] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 1695–1699.

[23] M. H. Rahman, I. Himawan, M. McLaren, C. Fookes, and S. Sridharan, “Employing phonetic information in DNN speaker embeddings to improve speaker recognition performance,” in *Proc. Interspeech*, 2018, pp. 3593–3597.

[24] J. Ma, V. Sethu, E. Ambikairajah, and K. A. Lee, “Speaker-phonetic vector estimation for short duration speaker verification,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5264–5268.

[25] T. Liu, R. K. Das, M. Madhavi, S. Shen, and H. Li, “Speaker-utterance dual attention for speaker and utterance verification,” in *Proc. Interspeech*, 2020, pp. 4293–4297.

[26] Y. Liu, Z. Li, L. Li, and Q. Hong, “Phoneme-aware and channel-wise attentive learning for text dependentspeaker verification,” in *Proc. Interspeech*, pp. 101–105, 2021.

[27] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[28] J. Zhong, W. Hu, F. K. Soong, and H. Meng, “DNN i-vector speaker verification with short, text-constrained test utterances,” in *Proc. Interspeech*, 2017, pp. 1507–1511.

[29] Z. Shi, H. Lin, L. Liu, and R. Liu, “Double joint Bayesian modeling of DNN local i-vector for text dependent speaker verification with random digit strings,” in *Proc. Interspeech*, 2018, pp. 67–71.

[30] S. Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[31] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4580–4584.

[32] Z. Wang, L. He, X. Gao, and Y. Huang, “Multi-scale spatial-temporal network for person re-identification,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2019, pp. 2052–2056.

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

[34] J.-W. Jung, H.-s. Heo, I.-L. Yang, H.-J. Shim, and H.-J. Yu, “Avoiding speaker overfitting in end-to-end DNNs using raw waveform for text-independent speaker verification,” in *Proc. Interspeech*, 2018, pp. 3583–3587.

[35] R. Liu, B. Sisman, G. lai Gao, and H. Li, “Expressive TTS training with frame and style reconstruction loss,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1806–1818, 2021, doi: 10.1109/TASLP.2021.3076369.

[36] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, “The Microsoft 2017 conversational speech recognition system,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2018, pp. 5934–5938.

[37] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, “Learning the speech front-end with raw waveform CLDNNs,” in *Proc. Interspeech*, 2015, pp. 1–5.

[38] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[39] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[41] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 1695–1699.

[42] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” *Soviet Phys. Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[43] T. Liu, M. C. Madhavi, R. K. Das, and H. Li, “A unified framework for speaker and utterance verification,” in *Proc. Interspeech*, 2019, pp. 4320–4324.