INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Speaker and Phoneme-Aware Speech Bandwidth Extension with Residual Dual-Path Network

Nana Hou[1], *Chenglin Xu*[1,4], *Van Tung Pham*[1], *Joey Tianyi Zhou*[3], *Eng Siong Chng*[1,2], *Haizhou Li*[4,5]

[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]Temasek Laboratories, Nanyang Technological University, Singapore
[3]Institute of High Performance Computing (IHPC), A*STAR, Singapore
[4]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[5]Machine Listening Lab, University of Bremen, Germany

nana001@e.ntu.edu.sg

Abstract

Speech bandwidth extension aims to generate a wideband signal from a narrowband (low-band) input by predicting the missing high-frequency components. It is believed that the general knowledge about the speaker and phonetic content strengthens the prediction. In this paper, we propose to augment the low-band acoustic features with i-vector and phonetic posteriorgram (PPG), which represent speaker and phonetic content of the speech, respectively. We also propose a residual dual-path network (RDPN) as the core module to process the augmented features, which fully utilizes the utterance-level temporal continuity information and avoids gradient vanishing. Experiments show that the proposed method achieves 20.2% and 7.0% relative improvements over the best baseline in terms of log-spectral distortion (LSD) and signal-to-noise ratio (SNR), respectively. Furthermore, our method is 16 times more compact than the best baseline in terms of the number of parameters.

Index Terms: Speech bandwidth extension, Residual dual-path network, Speaker and phoneme knowledge, I-vector, Phonetic posteriorgram

1. Introduction

Wideband speech signals are of higher perceptual quality and intelligibility. Speech bandwidth extension aims to predict the missing high-frequency components from narrowband (low-band) signals. It is an enabling technology for many applications, such as hearing aids [1], speech recognition and speaker verification [2–4]. Prior studies [5–10] are focused on estimating high-frequency magnitude and phase spectra in frequency domain. To overcome the inherent difficulty of phase estimation, time-domain frameworks [11–15] are proposed, that offer competitive voice quality.

However, speech bandwidth extension is still far from perfect due to many factors, such as unseen speaking voice or corrupted signals at run-time. Each speaker has unique formant structure [16]. As a result, bandwidth extension systems respond poorly to unseen speakers during inference. Besides, some low-band signals suffer from weak energy, such as consonants, which makes it even harder to recover the high-frequency components. There have been attempts to model the bandwidth extension in advance to alleviate these problems, for example, by extracting the speaker vocal tract shape information through a codebook search [17], or by using bottleneck features [7] to represent phoneme information in a recurrent neural network (RNN). The prior studies are the source of inspiration of this work.

Each speaker has a unique voice timbre and speech prosody. We hear clearer if we adapt our listening effort to the speaking voice, and the speech content. Implementing this idea, we propose to augment acoustic features with speaker and phonetic information in speech bandwidth extension. More specifically, we utilize i-vector, a low-dimensional speaker embedding [18–20], to characterize the voice identity of a speaker. We also use phonetic posteriorgram (PPG), which is a frame-based time-versus-class vector derived from automatic speech recognition system [21] to represent speech content. With concatenated i-vector and PPG features, we propose a speaker and phoneme-aware (SPA) speech bandwidth extension framework. As both i-vector and PPG come from the same input speech utterance, no additional information is required during run-time inference.

For effective prediction, we further adopt a residual dual-path network (RDPN) as the core module in the bandwidth extension network, which benefits from the utterance-level optimization and avoids gradient vanishing. Experiments show that RDPN outperforms the best baseline TFNet [22], and the speaker and phoneme-aware (SPA) RDPN system achieves a performance record with 16 times fewer parameters than TFNet.

This paper is organized as follows. In Section 2, we describe the proposed SPA-RDPN architecture. In Section 3, experimental settings and results are presented. Section 4 concludes the study.

2. SPA-RDPN Architecture

2.1. SPA-RDPN Network

We now introduce the proposed SPA-RDPN network, which consists of four modules: the encoder, SPA-features fusion, the core RDPN module and the decoder, as illustrated in Figure 1.

The convolutional encoder "conv", consisting of a 1-D convolutional layer, extracts acoustic features from the input speech, which is widely used in enhancement and separation task [23–25]. We employ 64 filters with a variety of filter sizes and a stride of half filter sizes. Such 64-dimensional acoustic features are then concatenated with the extracted i-vector and PPG features for the "feature fusion layer" to learn a low dimension encoding. Then, the core RDPN module is utilized to predict the missing high-frequency components from the fused features. Finally, a de-convolutional decoder "de-conv" reconstructs the speech waveform from the bandwidth-extended features. The details of SPA-features fusion and the core RDPN module are described as follows.

http://dx.doi.org/10.21437/Interspeech.2020-1994

2.2. Speaker, phonetic and acoustic feature fusion

I-vector is a low-dimensional feature vector, which characterizes a speaker. An i-vector extractor includes a universal background model (UBM) and total variability matrix. The extractor is trained with the MFCC features, together with energy plus their 1st and 2nd derivatives, which are extracted from narrowband speech of training and development sets of VCTK dataset [26]. Such features are then followed by cepstral mean normalization [27, 28]. We use a gender-independent UBM of 512 mixtures and a total variability matrix with 100 total factors.

PPG is a time-versus-class vector that represents the posterior probabilities of phonetic classes for a specific time frame. Unlike bottleneck features which is a type of phonetic embedding, PPG [29, 30] describes the phonetic classes defined by linguists. It can be obtained as the intermediate results of automatic speech recognition (ASR) decoding. We utilize a HMM-GMM based ASR system to derive a 37-dimension PPG feature for each speech frame with the forced alignment. The ASR system was trained using the default Kaldi scripts [21].

As discussed above, now we have encoded a speech utterance to a single 100-dimension i-vector, and a sequence of 37-dimension PPG features, each from one frame. We also encode the input utterance into a sequence of 64-dimension acoustic features. As the filter size and stride for acoustic feature is usually smaller than those of a PPG frame, the same utterance may generate more acoustic features than PPG frames. We therefore linearly upsample the PPG frames to frame-align with acoustic features. Finally we concatenate the acoustic feature, PPG feature, and i-vector to form an input to the feature fusion layer, which is a fully connected layer to project the 64+37+100 dimension vector to a lower dimension encoding.

2.3. RDPN core module

Though the feature dimension is reduced by the feature fusion layer, the length of features is still a huge number. Learning such a long feature sequence poses a challenge to conventional sequential modeling networks, including recurrent neural network and 1-D convolutional networks. Dilated convolutional layers were proposed to alleviate such problem [13, 14]. Despite the increased receptive field over the input signals, they have not captured the utterance-level temporal information.

To fully utilize the utterance-level temporal information, we propose the residual dual-path network (RDPN) as the core module, which consists of three stages: segmentation, residual block processing, and reconstruction, as illustrated in Figure 1 (in green).

Segmentation: As shown in Figure 1(i), we denote the feature fusion output as $X \in \mathbb{R}^{N \times L}$, where N is the feature dimension and L is the number of feature length. The operation of "segmentation" aims to segment the 2-D fused features X to form 3-D features, which splits X into M chunks of segments, each having a length of K with 50% overlap. The first and last chunks are zero-padded so that all speech frames in X are included. The chunks are then concatenated together to form a 3-D features (without counting in batch size) $X' \in \mathbb{R}^{N \times K \times M}$, where $M = \lceil 2L/K + 1 \rceil$ is the number of segments.

Residual block processing: As shown in Figure 1, the 3-D tensor X' is then taken by the stack of B residual dual-path blocks. Each block contains two sub-models named intra-chunk and inter-chunk processing, respectively. The intra-chunk and inter-chunk are of the same structure except that they process 3-D features along different dimensions, which is achieved by

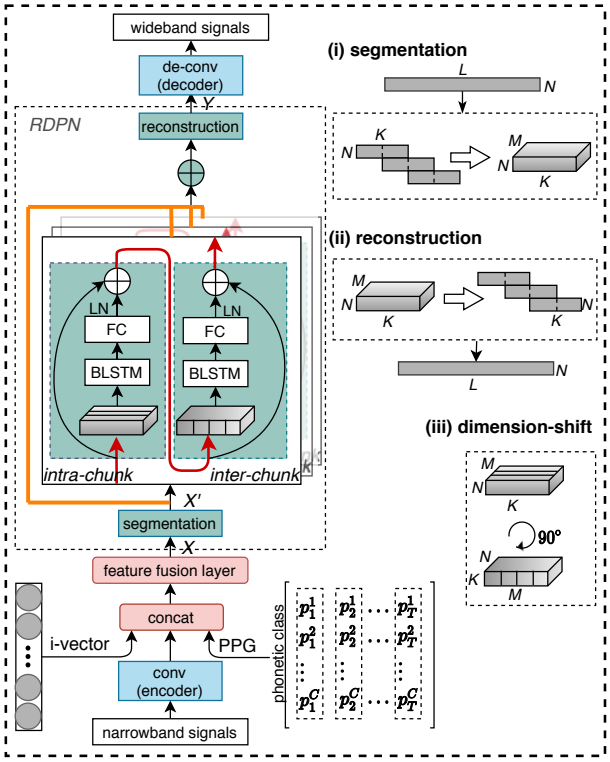

Figure 1: *The proposed SPA-RDPN network consists of 4 modules: the encoder, SPA-feature fusion, RDPN core module, and decoder. $\oplus$ denotes the residual connection. The narrowband and wideband signals are sampled at 16kHz, while the high-frequency components of narrowband signals are missing.*

"dimension-shift" as shown in Figure 1(iii).

The intra-chunk RNN consists of one bi-directional long-short-term-memory (BLSTM) layer, one linear fully-connected layer (FC) and a layer normalization [31]. The BLSTM layer in intra-chunk is applied to the second dimension (i.e., K) of $X' \in \mathbb{R}^{N \times K \times M}$ to learn local information in every segmentation of X'. Conversely, the BLSTM in inter-chunk performs along the third dimension (i.e., M) of features to learn global information across every segmentation. This operation is easily achieved by "dimension-shift" before they are fed into the BLSTM layer. The FC layer is used for feature dimension consistence.

We perform the following layer normalization (LN) for the module to have a good generalization ability:

$$\mathrm{LN}(X'') = \frac{X'' - \mu(X'')}{\sqrt{\sigma(X'') + \epsilon}} \odot \alpha + \beta \qquad (1)$$

where X'' is the output from FC layer, μ and σ are the mean and variance of X'', and $\alpha, \beta \in \mathbb{R}^{N \times 1}$ are the rescaling factors. The orange residual connections in Figure 1 allow the flow of memory (or information) from initial blocks to last blocks so as to avoid the gradient vanishing.

Reconstruction: As shown in Figure 1(ii), "reconstruction" is the inverse operation of "segmentation", which aims to flat out the 3-D bandwidth-extended embedding from the last residual dual-path block back to 2-D feature sequence Y. The dimension of Y is consistent with 2-D narrowband feature $X \in \mathbb{R}^{N \times L}$. Afterwards, the bandwidth-extended features $Y \in \mathbb{R}^{N \times L}$ is recovered to speech waveforms by the 1-D deconvolutional decoder.

3. Experiments and Results

3.1. Dataset

We evaluated our experiments on VCTK dataset[1], which contained 44 hours speech data from 109 native English speakers [26]. Following the previous works [11, 22], we divided the VCTK dataset into a training set (an amount of 88% data), a validation set (an amount of 6% data), and a test set (an amount of 6% data)[2]. The speakers in the test set were unseen during the training stage. To prepare the training pairs, we sampled both narrowband and wideband speech signals at 16kHz, and the narrowband speech was pre-processed by Bicubic algorithm like other work [11, 22].

3.2. Experimental Setup

3.2.1. Network configuration

The encoder consists of a 1-D convolutional layer with 64 filters followed by a rectified linear unit (ReLU) activation function. We tuned the window length of the filters using various settings $w(= 2, 4, 8, 16$ samples) with a stride of $w/2$ samples. The "feature fusion layer" is a non-linear dense layer with 128 nodes and a ReLU activation function. We empirically set $B(= 6)$ residual dual-path blocks, where the BLSTM layer is with 128 hidden units in each direction. The following FC layer is utilized to keep the dimension consistence with input of residual dual-path blocks. We utilize the scale-invariant signal-to-distortion ratio (SI-SDR) [32] as the loss function.

The network was optimized by the Adam algorithm [33]. The learning rate started from 0.0001 and was halved when the loss increased on the development set for at least 3 epochs. Early stopping scheme was applied when the loss increased on the development set for 20 epochs. The utterances in the training and development set were broken into 16,384 samples segments for batch training, and the minibatch size was set to 16. The UBM-GMM based i-vector extractor and the HMM-GMM based ASR system were trained using narrowband speech from VCTK by Kaldi default scripts [21].

3.2.2. Reference baselines

We compare our results with 4 reference baselines that represent the recent advancements in speech bandwidth extension.

- **LSM** [5]: applied 3 feed forward layers to predict the missing high-frequency components from the low-frequency log-spectrum in frequency domain. The missing high-frequency phase was recovered by imaged phase of low-frequency signals.
- **DRCNN** [11]: mapped narrowband audios to wideband via fully convolutional encoder-decoder framework in the time domain. To increase the time dimensions during upscaling, subpixel shuffling layers were introduced in the upsampling blocks. The skip connections were utilized to speed up training.
- **WaveNet** [14]: utilized wavenet based vocoder, consisting of stacked dilated convolutional layers conditioned on a log-mel spectrum representation, to reconstruct the missing high-frequency components.
- **TFNet** [22]: utilized supervision in both the frequency- and time-domain to jointly optimize the previous architecture [11]. A spectrum fusion layer was proposed to combine the features from frequency- and time-domain with a learnable parameter.

3.3. Results

We report the performance in terms of log-spectral distortion (LSD) [34], signal-to-noise ratio (SNR), signal-to-distortion ratio (SDR) [35] and perceptual evaluation of the speech quality (PESQ) [36]. Except LSD, higher scores are better for all metrics. The subjective evaluation of A/B preference test was also conducted.

3.3.1. Effect of window length on system performance

We first analyse and summarize the system performances with different window length of encoder and decoder filters. I-vector and PPG features are not utilized in this experiment. As shown in Figure 2, the x-axis represents the window length $(= 2, 4, 8, 16$ samples or $0.125, 0.25, 0.5, 1$ millisecond) of encoder and decoder filters of the system. The four y-axis are the evaluation metrics in terms of LSD, SNR, SDR, and PESQ, respectively. We observe that the performances consistently degrade by further increasing the window length in encoder and decoder filters. The best performance is obtained when the filter length is 2 samples with an encoder output of more than 16,000 frames. Such long-range temporal continuity information can be extremely hard or even impossible for standard RNNs or CNNs to model, while with the proposed RDPN, the usage of such a short window becomes possible and achieves the best performance. The following experiments adopt this setting.

3.3.2. Effect of i-vector and PPG

We further summarize the performances by integrating i-vector and PPG into the proposed RDPN in Table 1. We observe that the performances are gradually improved with speaker and phonetic knowledge. Compared with the RDPN, SPA-RDPN achieves 11.8% and 5.0% relative improvements in terms of LSD and SNR with approximate same number of parameters. We also observe that phonetic information brings higher improvement than speaker characteristics. The quality of the recovered wideband signal (e.g., PESQ) has been significantly improved by utilizing the speaker and phonetic information.

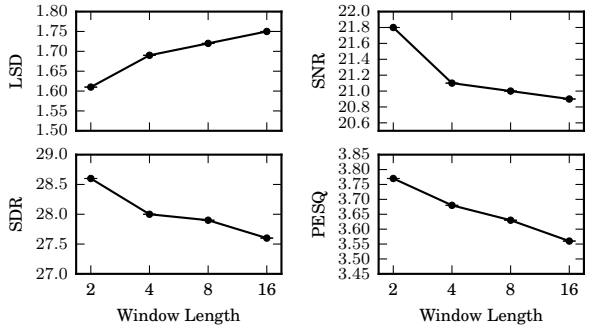

Figure 2: *LSD, SNR(dB), SDR(dB) and PESQ in a comparative study of different window lengths of 1-D convolutional filters of the encoder and decoder in the proposed system.*

[1]The dataset is available at: `https://datashare.is.ed.ac.uk/handle/10283/2651`

[2]The split testset is available at: `https://github.com/nanahou/spa-rdpn/blob/master/test_speakers`

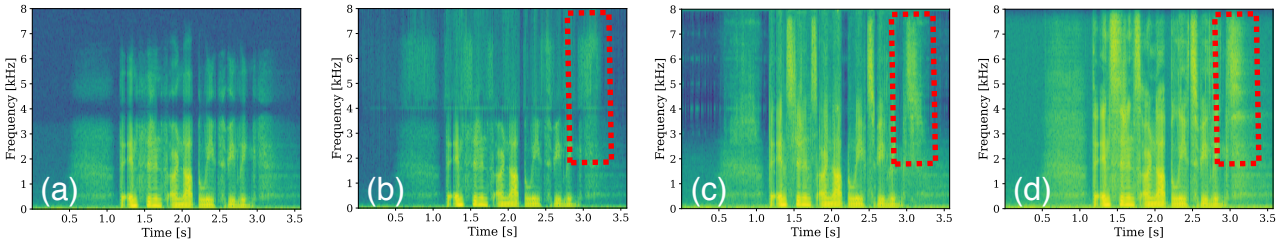

Figure 3: *The spectrogram illustrations of the example (p360_059.wav) in the testing set of VCTK about (a) narrowband input pre-processed by Bicubic algorithm, (b) the best baseline TFNet, (c) SPA-RDPN and (d) wideband (ground-truth).*

Table 1: *LSD, SNR(dB), SDR(dB) and PESQ in a comparative study of the proposed SPA-RDPN. "#Paras" means the number of parameters of the model.*

Methods	#Paras	LSD	SNR	SDR	PESQ
RDPN	3.59M	1.61	21.8	28.6	3.77
RDPN+i-vector	3.61M	1.55	22.1	29.0	3.85
SPA-RDPN	3.62M	**1.42**	**22.9**	**29.7**	**3.98**

Table 2: *LSD and SNR(dB) in a comparative study of recent state-of-the-art techniques. "Feature type" refers to the types of narrowband inputs. "Spectrum" means that the approach is performed in frequency domain and "waveform" means that the time-domain signals are directly taken as inputs. "Mixed" refer to utilizing both frequency-domain and time-domain features. The TFNet [22] was re-implemented by the paper on VCTK dataset.*

Methods	#Paras	Feature type	LSD	SNR
DNN [5]	13.38M	spectrum	3.60	19.9
DRCNN [11]	56.41M	waveform	3.10	20.7
WaveNet [14]	15.13M	waveform	2.31	–
TFNet [22]*	58.18M	mixed	1.78	21.4
SPA-RDPN	**3.62M**	waveform	**1.42**	**22.9**

3.3.3. Overall comparison

Table 2 summarizes the comparison between the proposed SPA-RDPN and other recent state-of-the-art techniques in terms of LSD and SNR. We observe that the proposed SPA-RDPN obtained the best performance. Comparing with the TFNet method, the SPA-RDPN achieves 20.2% and 7.0% relative improvements in terms of LSD and SNR. Meanwhile, the parameter size of the SPA-RDPN is 16 times smaller than that of the TFNet.

To further show the contribution of the proposed SPA-RDPN, we illustrate the magnitude spectrum of an example as shown in Figure 3. We can see despite the energy of some speech parts in low-frequency is weak, the proposed SPA-RDPN still can recover the richer high-frequency information than the best baseline TFNet.

3.3.4. Subjective evaluation

Since the TFNet presents the best baseline performance in the objective evaluation as shown in Table 2, we only conduct an A/B preference test between the TFNet and the proposed SPA-RDPN to evaluate the signal quality and intelligibility by subject listening. We randomly selected 20 pairs of listening examples and invited 10 subjects to choose their preference according to the quality and intelligibility of the recovered wideband signal. The percentage of the preferences is shown in Figure 4. We observe that the listeners clearly preferred the proposed SPA-RDPN with a preference score of 74% to the best baseline TFNet with a preference score of 23%. Most subjects significantly preferred the reconstructed wideband signals by SPA-RDPN with a significance level of $p < 0.05$, because the audios sound more natural and have better quality. Some listening examples are available at Github[3].

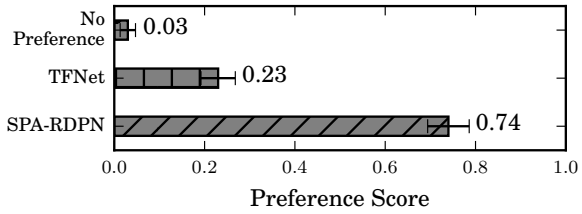

Figure 4: *The A/B preference test result of the recovered speech between the proposed SPA-RDPN and the best baseline TFNet. We conducted t-test using a significance level of $p < 0.05$, which is depicted with error bars.*

4. Conclusions

In this paper, we proposed a speaker and phoneme-aware speech bandwidth extension approach by incorporating i-vector and phonetic posteriorgram (PPG) features into the residual dual-path network (SPA-RDPN). Experimental results show that the proposed SPA-RDPN outperforms the best baseline TFNet in terms of LSD and SNR with 16 times fewer parameters. The subjective test also shows that the SPA-RDPN is significantly preferred comparing with the TFNet.

5. Acknowledgements

This work was supported by Air Traffic Management Research Institute of Nanyang Technological University, Human-Robot Interaction Phase 1 (Grant No. 192 25 00054), National Research Foundation (NRF) Singapore under the National Robotics Programme; AI Speech Lab (Award No. AISG-100E-2018-006), NRF Singapore under the AI Singapore Programme; Human Robot Collaborative AI for AME (Grant No. A18A2b0046), NRF Singapore; Neuromorphic Computing Programme (Grant No. A1687b0033), RIE 2020 AME Programmatic Grant. The work by H. Li is also partly supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany).

[3] https://nanahou.github.io/spa-rdpn/

6. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[2] K. Li, Z. Huang, Y. Xu, and C.-H. Lee, "Dnn-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrow-band speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[3] R. Kaminishi, H. Miyamoto, S. Shiota, and H. Kiya, "Investigation on blind bandwidth extension with a non-linear function and its evaluation of x-vector-based speaker verification," in *Proc. INTERSPEECH*, 2019, pp. 4055–4059.

[4] H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for deep speaker embedding," *Proc. Interspeech 2019*, pp. 406–410, 2019.

[5] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4395–4399.

[6] B. Liu, J. Tao, Z. Wen, Y. Li, and D. Bukhari, "A novel method of artificial bandwidth extension using deep architecture," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] Y. Gu, Z.-H. Ling, and L.-R. Dai, "Speech bandwidth extension using bottleneck features and deep recurrent neural networks." in *Interspeech*, 2016, pp. 297–301.

[8] J. Abel, M. Strake, and T. Fingscheidt, "A simple cepstral domain dnn approach to artificial speech bandwidth extension," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5469–5473.

[9] S. E. Eskimez and K. Koishida, "Speech super resolution generative adversarial network," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3717–3721.

[10] P. Bachhav, M. Todisco, and N. Evans, "Latent representation learning for artificial bandwidth extension using a conditional variational auto-encoder," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7010–7014.

[11] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super-resolution using neural nets," in *ICLR (Workshop Track)*, 2017.

[12] M. Wang, Z. Wu, S. Kang, X. Wu, J. Jia, D. Su, D. Yu, and H. Meng, "Speech super-resolution using parallel wavenet," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 260–264.

[13] Z.-H. Ling, Y. Ai, Y. Gu, and L.-R. Dai, "Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 883–894, 2018.

[14] A. Gupta, B. Shillingford, Y. Assael, and T. C. Walters, "Speech bandwidth extension with wavenet," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 205–208.

[15] X. Hao, C. Xu, N. Hou, L. Xie, E. S. Chng, and H. Li, "Time-domain neural network approach for speech bandwidth extension," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 866–870.

[16] L. R. Rabiner and B. Gold, "Theory and application of digital signal processing," *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975. 777 p.*, 1975.

[17] I. Katsir, I. Cohen, and D. Malah, "Speech bandwidth extension based on speech phonetic content and speaker vocal tract shape estimation," in *2011 19th European Signal Processing Conference*. IEEE, 2011, pp. 461–465.

[18] W. Rao, M.-W. Mak, and K.-A. Lee, "Normalization of total variability matrix for i-vector/plda speaker verification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4180–4184.

[19] W. Rao, C. Xu, E. S. Chng, and H. Li, "Target speaker extraction for multi-talker speaker verification," *Proc. Interspeech 2019*, pp. 1273–1277, 2019.

[20] C. Xu, W. Rao, E. S. Chng, and H. Li, "Time-domain speaker extraction network," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 327–334.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[22] T. Y. Lim, R. A. Yeh, Y. Xu, M. N. Do, and M. Hasegawa-Johnson, "Time-frequency networks for audio super-resolution," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 646–650.

[23] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[24] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," *arXiv preprint arXiv:1910.06379*, 2019.

[25] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[26] J. Yamagishi, "English multi-speaker corpus for cstr voice cloning toolkit," *URL http://homepages. inf. ed. ac. uk/jyamagis/page3/page58/page58. html*, 2012.

[27] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[28] N. Hou, C. Xu, E. S. Chng, and H. Li, "Domain adversarial training for speech enhancement," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 667–672.

[29] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6790–6794.

[30] X. Tian, J. Wang, H. Xu, E. S. Chng, and H. Li, "Average modeling approach to voice conversion with non-parallel data." in *Odyssey*, 2018, pp. 227–232.

[31] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[32] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[34] L. Rabiner, "Fundamentals of speech recognition," *Fundamentals of speech recognition*, 1993.

[35] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[36] I. Rec, "P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union*, 2005.