INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Waveform Modeling Using Stacked Dilated Convolutional Neural Networks for Speech Bandwidth Extension

Yu Gu, Zhen-Hua Ling

National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China, Hefei, P.R.China
hicolin@mail.ustc.edu.cn, zhling@ustc.edu.cn

Abstract

This paper presents a waveform modeling and generation method for speech bandwidth extension (BWE) using stacked dilated convolutional neural networks (CNNs) with causal or non-causal convolutional layers. Such dilated CNNs describe the predictive distribution for each wideband or high-frequency speech sample conditioned on the input narrowband speech samples. Distinguished from conventional frame-based BWE approaches, the proposed methods can model the speech waveforms directly and therefore avert the spectral conversion and phase estimation problems. Experimental results prove that the BWE methods proposed in this paper can achieve better performance than the state-of-the-art frame-based approach utilizing recurrent neural networks (RNNs) incorporating long short-term memory (LSTM) cells in subjective preference tests.

Index Terms: speech bandwidth extension, stacked dilated convolutional neural networks, causal convolution, non-causal convolution, WaveNet

1. Introduction

Due to the restriction of speech acquisition equipments and transmission systems, the bandwidth of speech signal is usually limited to a particular narrowband of frequencies. Although the intelligibility of narrow speech is satisfactory, the absence of high-frequency counterpart leads to a muffled sound, resulting in seriously degraded speech quality, naturalness and speaker-similarity. Speech bandwidth extension (BWE) techniques aim at automatically restoring the missing high-frequency components of narrowband speech by exploiting the correlation that exists between low and high frequency parts of wideband speech. A well-built BWE system can not only bring in a dramatic improvement of perceived speech quality for conventional telephone networks but also benefit other speech processing tasks such as speech enhancement [1] and recognition [2].

BWE algorithms have been studied for decades and a large amount of methods have been proposed to further improve the quality of narrowband speech. There were some simple methods such as codebook mapping [3], linear mapping [4] and rule-based spectrum folding, and some more complicated statistical approaches using Gaussian mixture models (GMMs) [5, 6, 7] and hidden Markov models (HMMs) [8, 9, 10]. Nevertheless, these methods suffer from the over-smoothing effect and severe artifacts due to their deficient ability of acoustic modeling.

This work was partially funded by the CAS Strategic Priority Research Program (Grant No. XDB02070006), the Fundamental Research Funds for the Central Universities (Grant No. WK2350000001), the National Key Research and Development Program (Grant No. 2016YFB1001300) and National Natural Science Foundation of China (Grant No. U1613211).

Deep learning technology has been intensively studied and explored by speech signal processing researchers in recent years. In many speech generation tasks such as voice conversion, speech enhancement, articulatory-to-acoustic mapping and text-to-speech synthesis [11], different kinds of neural networks with various deep structures have shown remarkable acoustic modeling capabilities and better performances than conventional methods based on GMMs or HMMs. Several stochastic neural networks such as restricted Boltzmann machines, bidirectional associative memories [12] and DNNs with different structures and training strategies [2, 12, 13, 14, 15, 16] have also been adopted in BWE tasks to replace GMMs or HMMs to model the sophisticated and non-linear mapping relationship from narrowband speech parameters to high-frequency ones. Because of their better ability of modeling high-dimensional observations with cross-dimension correlations, raw and high-dimensional spectral envelopes or magnitude spectra rather than their low-dimensional representations can be utilized directly in those DNN-based methods. Deep unidirectional or bidirectional RNNs with stacked layers of LSTM cells [17, 18] have also been adopted to model the temporal dependencies among the sequences of low-frequency and high-frequency spectral features. RNNs can effectively alleviate the discontinuity caused by the frame-independent mapping functions in feed-forward DNNs. The experimental results of previous work [12, 13, 14, 15] have proved that DNN-based BWE methods can effectively alleviate the over-smoothing effect and improve the speech quality of BWE outputs compared with conventional GMM-based ones. Meanwhile RNN-based methods [17, 18] can further improve the BWE reconstruction accuracy and acquire better subjective listening performances.

Existing BWE approaches are usually frame-based and the core procedures of converting input narrowband speech to wideband output are conducted in frequency domain. Current BWE methods also mainly concentrate on addressing the problems of modeling the intrinsic correlation or mapping relationship of the magnitude spectra. The model-based prediction of the high-frequency phase spectra is always difficult due to the issue of phase wrapping. However, the inaccuracy of the reconstructed high-frequency phase spectra can also lead to severe whistling and hissing artifacts and metallic sounds.

Convolutional neural networks have enjoyed great popularity as means for image processing and have also been employed as acoustic models in speech recognition tasks [19, 20, 21, 22]. Convolutional layers can aggregate feature extractors by directly operating on raw signal such as image pixels and speech waveforms. In addition to classification tasks, various kinds of CNNs, such as WaveNet [23] for text-to-speech synthesis and ByteNet [24] for machine translation, have achieved significant improvement on generation tasks. Therefore motivated by the success of the dilated convolution architectures [25, 26] as

http://dx.doi.org/10.21437/Interspeech.2017-336

well as WaveNet and ByteNet, a BWE method using stacked dilated CNNs is present in this paper to avoid the spectral analysis and phase modeling issues by directly modeling and generating speech waveforms.

The rest of this paper is organized as follows. Section 2 gives a brief review of the conventional frame-based BWE approaches using deep structured neural networks. Section 3 describes the stacked dilated CNNs and the structures of residual and skip connections applied in this paper. Our proposed methods are introduced in detail in Section 4. Section 5 shows the experimental results and Section 6 concludes this paper.

2. Frame-based BWE using magnitude spectra and regression neural networks

Almost all the current BWE systems are frame-based with certain frame length and frame shift. A vocoder is an indispensable part for feature extraction and the speech signal transformation between time and frequency domains. Logarithmic magnitude spectrum derived from the short-time Fourier transform [27] on speech waveforms is one of the most popular frame-based acoustic features for BWE tasks. Various full-connection regression neural networks such as DNNs and RNNs with stacked LSTM cells have been employed to estimate the mapping function from narrowband magnitude spectra to their high-frequency counterparts under minimum mean squared error criterion [13, 15, 17, 18]. At the stage of restoration, the log power spectra of wideband speech were reconstructed by concatenating the input magnitude spectra of narrowband speech and the high-frequency magnitude spectra predicted using the trained networks. The phase spectra of wideband speech were usually estimated from the phase spectra of narrowband speech by mirror inversion [13, 15]. Finally, inverse FFT and overlap-add algorithm were carried on to reconstruct the wideband waveforms according to the extended magnitude and phase spectra.

3. Dilated CNNs for waveform generation

3.1. Dilated CNNs

Convolutional neural networks are a specialized kind of feed-forward neural networks for signal processing, which provide shift-invariance and weight sharing properties over time or space. Such neural networks have been tremendously successful in practical applications such as image recognition. CNNs can also be used as sequence models with a fixed dependency range by employing a mathematical convolution operation on time-series data. It has been demonstrated that CNNs can further reduce word error rate by modeling directly on raw speech waveforms in speech recognition [20, 21, 22]. As for speech generation tasks, DeepMind's WaveNet [23] was proposed for text-to-speech synthesis and other general audio generation tasks, which was capable of producing significantly more natural sounds than conventional approaches. Different from existing paradigms for parametric speech generation algorithms, WaveNet performed autoregressive speech sample generation using an acoustic model with dilated causal convolutional layers instead of depending on vocoders. The architectures of exploited dilated CNNs are illustrated on Figure 1. The causal convolutional layers have various dilation factors that allow their receptive field to grow exponentially in terms of the depths of networks as opposed to linearly, and can therefore cover the input history information from thousands of timesteps ahead. For obtaining a large range of receptive field, an extremely deep structure with many convolutional layers is always needed. Such deep dilated CNN can be constructed by stacking multiple convolutional layers one on the top of another, and the output sequence of lower layer is considered as the input sequence for the following layer. The dilated causal CNN can be regarded as a statistical model and the conditional distribution of the output sample sequence $\boldsymbol{y} = \{y_1, y_2, \cdots, y_T\}$ given the input sequence $\boldsymbol{x} = \{x_1, x_2, \cdots, x_T\}$ is factorized as the product of conditional probabilities as follows:

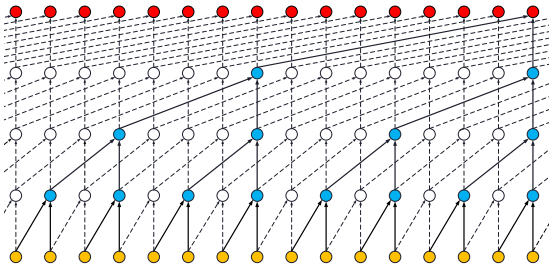

Figure 1: *Network structures of stacked dilated causal CNNs.*

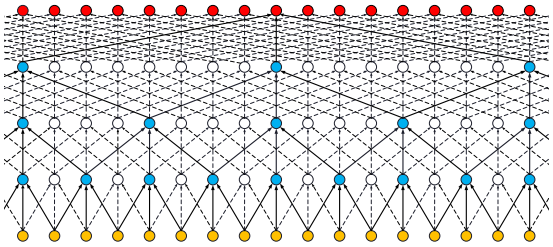

Figure 2: *Structures of stacked dilated non-causal CNNs.*

$$p\left(\boldsymbol{y} \mid \boldsymbol{x}\right) = \prod_{i=1}^{T} p\left(y_i \mid x_{i-N+1}, x_{i-N+2}, \cdots, x_i\right), \quad (1)$$

where N is the length of the receptive field. Such causal convolution structures are designed to capture a long range of past inputs, which can guarantee low latency and autoregressive generation mechanism as WaveNet.

Nevertheless, the future input information is also quite essential for the reconstruction of the output sequence. The dilated non-causal convolution architectures like DeepMind's ByteNet [24] as depicted on Figure 2 are able to take full advantages of the context of the input sequence. Hence such non-causal layout is especially efficient when subsequent input information is required or necessary. Then Equation (1) should be rewritten as following:

$$p\left(\boldsymbol{y} \mid \boldsymbol{x}\right) = \prod_{i=1}^{T} p\left(y_i \mid x_{i-N/2}, x_{i-N/2+1}, \cdots, x_{i+N/2}\right), \quad (2)$$

where $N+1$ is the length of corresponding receptive field.

3.2. Residual blocks and gated activation units

Neural networks with many hidden layers usually suffer from the issues of training accuracy degradation and slow convergence. They can't be easily optimized as shallower networks. Residual learning strategies [28] were invented to address these issues. A variant of residual blocks was also applied on the dilated CNNs in WaveNet. As exhibited in Figure 3, each convolutional layer is wrapped in such a residual block which contains gated activation units and two additional convolutional layers

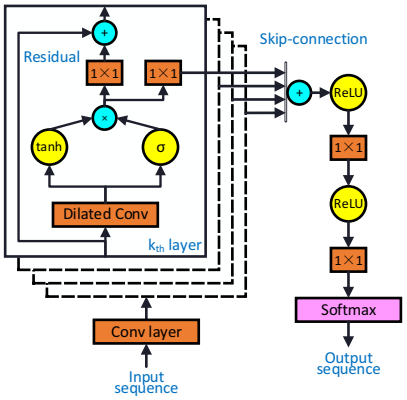

Figure 3: *Diagrams of the residual blocks.*

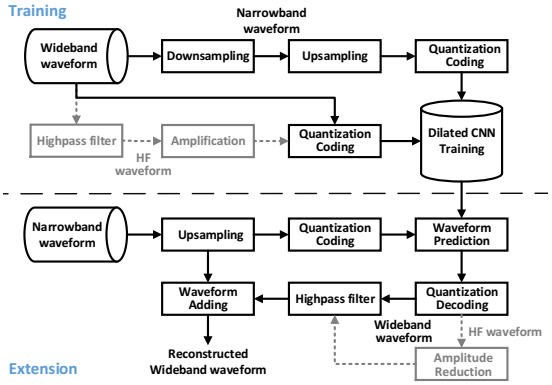

Figure 4: *Flowchart of the proposed BWE approaches.*

towards the following and output layers respectively with convolution filters of size 1. The residual and parameterized skip-connections are deployed throughout the network to capacitate training deeper networks and to accelerate convergence. The gated activation units in k-th layer are expressed as:

$$\hat{\boldsymbol{h}}_k = \tanh(\boldsymbol{W}_{f,k} * \boldsymbol{h}_k) \odot \sigma(\boldsymbol{W}_{g,k} * \boldsymbol{h}_k), \qquad (3)$$

where f and g denote the filter and gate parts respectively, σ is the sigmoid non-linearity function, $\odot$ is the element-wise product and $*$ is the convolution operator. The output layer is cascaded with a softmax layer and thus the model can describe the categorical distribution over the output sequence.

4. BWE using dilated CNNs

Our proposed BWE approaches using dilated CNNs follow the framework illustrated in Figure 4. Comparing with the conventional BWE approaches introduced in Section 2, the proposed BWE method can omit all procedures related with vocoders, such as feature extraction and waveform reconstruction from the restored wideband features. The method is performed directly on the narrowband speech waveforms and their correlative wideband or high-frequency speech waveforms.

At the training stage, the parallel narrowband speech waveforms can be generated by down-sampling the wideband speech in the training corpus and are treated as network inputs. Unlike the output setting for standard WaveNet training, which was just the time-shift result of the input natural samples, the output sequences in our BWE methods are the waveform samples of wideband speech or high-frequency component. To guarantee the length consistency between the input and output sequences for model training, the narrowband speech should be up-sampled to the equal sampling rate with the wideband speech through zero-interpolation operation and a lowpass filter. Then the processed narrowband speech acts as the input for the acoustic models. Similar with WaveNet, all the input and output sequences are quantized to discrete values using μ-law [29] and one-hot coding is pursued on the quantized waveforms. Especially for the situation that the model outputs are the high-frequency waveforms, an extra procedure of amplification is appended to reduce the quantization error as shown by the grey dashed lines in Figure 4.

Both deep causal and non-causal CNNs with multiple dilated convolutional layers are utilized in this paper to model the temporal mapping relationship from the input narrowband sample sequence toward the output target sample sequence as shown in Figure 4. The low latency characteristic of causal convolution architecture is quite suitable for the real-time BWE applications such as speech communication and the non-causal structure can be competent for off-line BWE tasks. The network training is based on cross-entropy criterion to iteratively improve the classification accuracy of the network outputs with the target output sample sequences in training set. At the stage of extension, the target wideband or high-frequency waveforms can be generated sequentially from the sequence of input narrowband speech. Each output sample is drawn by selecting the quantization level with maximum posterior probability computed by the trained network. The generated waveforms are further processed by a highpass filter and are added with the input narrowband speech to reconstruct the final wideband speech.

5. Experiments

Our experiments adopted the TIMIT corpus [30], which contained English speech from multi-speakers sampled at 16kHz with 16-bit resolution. Parallel narrowband speech at 8kHz was produced by down-sampling the wideband speech at 16kHz in our experiments. 3696 utterances were chosen to construct the training set. 192 utterances were chosen as the validation set for model selection and another 192 utterances from the speakers not included in the training set were used as the test set to measure the performance of different BWE systems. The following five systems were established for comparison.

- ***RNN***: The frame-based method as introduced in Section 2 using RNN and logarithmic magnitude spectra [17];
- ***CNN1-WB***: The proposed causal CNN-based method, with wideband speech waveforms as output;
- ***CNN2-WB***: The proposed non-causal CNN-based method, with wideband speech waveforms as output;
- ***CNN1-HF***: The proposed causal CNN-based method, with high-frequency speech waveforms as output;
- ***CNN2-HF***: The proposed non-causal CNN-based method, with high-frequency waveforms as output.

5.1. Prediction accuracy validation

Figure 5 exhibits the prediction accuracies of the output speech samples on the validation set for the four proposed systems and the effect of using different receptive field lengths for the stacked dilated CNNs was also evaluated in our experiments. Comparing the ***CNN1-WB***, ***CNN1-HF*** systems with the ***CNN2-WB***, ***CNN2-HF*** systems, the experimental results demonstrate that the methods with non-causal convolutional layers achieved better prediction accuracies than those based on causal CNNs when the network outputs were either wideband

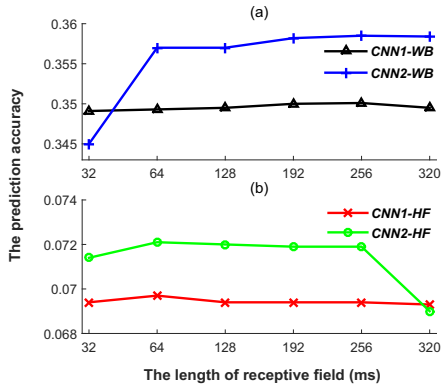

Figure 5: *Prediction accuracies on the validation set of the systems with different lengths of receptive field.*

speech waveforms or high-frequency speech waveforms, which demonstrates the effectiveness of importing the future context of input. The prediction errors for the ***CNN1-HF***, ***CNN2-HF*** systems were much larger than the ***CNN1-WB***, ***CNN2-WB*** systems for the reason that the noise-like and aperiodic high frequency waveforms are intrinsically much more intractable to accurately restore. Diverse ranges of the receptive field lengths from 32ms to 320ms were investigated as depicted in Figure 5. A wide scope of receptive field can help reduce the error of predicting wideband speech waveforms in the ***CNN1-WB***, ***CNN2-WB*** systems. However, for the ***CNN1-HF*** and ***CNN2-HF*** approaches, the accurate rates did not rise as increasing the length of receptive field, which indicates a short range of receptive field is sufficient for the situations when the network outputs are high-frequency waveforms. The aperiodic components are much stronger in high-frequency waveforms and the periodic properties are more remarkable in wideband speech. Therefore, a wide receptive field of multiple F0 periods are favorable in the ***CNN1-WB*** and ***CNN2-WB*** systems and a short receptive field is acceptable in ***CNN1-HF*** and ***CNN2-HF***. According to the results in Figure 5, the network configurations for each system were determined and are summarized in Table 1.

Table 1: *Network configurations for the CNN-based systems. n, f, r, l, rc and sc represent the number of layers, maximum dilated factor, receptive field length, theoretical latency, residual channel size and skip-connection channel size respectively.*

System	n	f	r (ms)	l (ms)	rc	sc
CNN1-WB	40	512	256	-	100	512
CNN2-WB	36	256	256	128	100	512
CNN1-HF	18	256	64	-	100	512
CNN2-HF	16	128	64	32	100	512

The non-causal convolution architectures must bring in a certain latency for catching sight of future input information, however they can possess identical receptive field with less convolutional layers and drive down computation costs. By using TensorFlow framework [31] and a single Tesla K40 GPU, a runtime of about five times slower than real-time is needed. However, the speed can be further accelerated by removing redundant computation and adopting pipeline strategies.

5.2. Subjective evaluation

Several preference tests were performed to assess the subjective perceptual quality of the extended wideband speech generated using different BWE systems. In each preference test, the wideband speech of 20 test utterances randomly selected from the test set were reconstructed by two different systems and evaluated in random order by six listeners. The listeners were asked to choose their preference for each given pairwise utterances in terms of speech quality.[1] The preference scores of these listening tests are exhibited in Figure 6 with the p-values from t-test. The comparisons of the conventional frame-based ***RNN*** system and the proposed the ***CNN1-WB***, ***CNN2-WB*** systems demonstrate the proposed method using waveform modeling and causal or non-causal dilated CNNs can successfully improve the quality of generated speech. The improvement for the ***CNN2-WB*** is more significant than the ***CNN1-WB*** system, which is consistent with the prediction accuracy difference between these two systems in Figure 5. The superiority of the ***CNN1-HF*** and ***CNN2-HF*** systems over the ***CNN1-WB*** and ***CNN2-WB*** systems on preference scores indicates the effectiveness of setting high-frequency waveforms as the network outputs rather than wideband waveforms. Although directly reconstructing the high-frequency waveforms is more challenging and the absolute values of sample prediction accuracy are much lower, it can reduce the redundancy of waveform prediction since the low-frequency component is known. The proposed ***CNN2-HF*** system which took advantages of both non-causal convolution structures and the strategy of directly reconstructing high-frequency waveforms also achieved better speech quality than the ***CNN1-HF*** system.

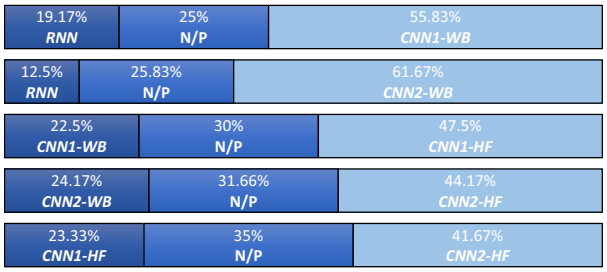

Figure 6: *Preference test scores among different BWE systems. The p-values of t-test in these comparisons are 1.3×10^{-6}, 8.1×10^{-12}, 8.7×10^{-4}, 7.5×10^{-3} and 0.012 respectively.*

6. Conclusions

In this paper, we have proposed a speech bandwidth extension method with waveform modeling and generation using dilated CNNs. Compared with the conventional frame-based BWE method using RNN and magnitude spectra, the proposed methods are better at modeling the temporal mapping relationship from narrowband input speech to the corresponding high-frequency component. Subjective experimental results show that the proposed methods obtained better preference scores than the RNN-based approach. Meanwhile the systems using non-causal convolution structures achieved better prediction accuracies. The methods using high-frequency waveforms as model outputs outperformed those using wideband waveforms as outputs. In our future work, further analysis and optimization of the model structures will be conducted and the conditional dilated CNNs with auxiliary condition information such as bottleneck features and acoustic features will also be investigated.

[1]Examples of restored wideband speech are available at `http://home.ustc.edu.cn/~hicolin/demos_IS2017.html`.

7. References

[1] Y. He, J. Han, T. Zheng, and G. Sun, "A new framework for robust speech recognition in complex channel environments," *Digital Signal Processing*, vol. 32, pp. 109–123, 2014.

[2] K. Li, Z. Huang, Y. Xu, and C. Lee, "DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech," in *INTERSPEECH 2015*, September 2015, pp. 2578–2582.

[3] S. Vaseghi, E. Zavarehei, and Q. Yan, "Speech bandwidth extension: Extrapolations of spectral envelop and harmonicity quality of excitation," in *Acoustics, Speech and Signal Processing (ICASSP), 2006 IEEE International Conference on*, vol. 3, May 2006, pp. III–III.

[4] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies," in *Acoustics, Speech, and Signal Processing (ICASSP), 2001 IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 665–668.

[5] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Acoustics, Speech, and Signal Processing (ICASSP), 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1843–1846.

[6] Y. Ohtani, M. Tamura, M. Morita, and M. Akamine, "GMM-based bandwidth extension using sub-band basis spectrum model," in *INTERSPEECH 2014*, September 2014, pp. 2489–2493.

[7] W. Fujitsuru, H. Sekimoto, T. Toda, H. Saruwatari, and K. Shikano, "Bandwidth extension of cellular phone speech based on maximum likelihood estimation with GMM," in *2008 RISP International Workshop on Nonlinear Circuits and Signal Processing*, 2008, pp. 283–286.

[8] G. Chen and V. Parsa, "HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies," in *Acoustics, Speech and Signal Processing (ICASSP), 2004 IEEE International Conference on*, vol. 1, May 2004, pp. I–709–12 vol.1.

[9] G.-B. Song and P. Martynovich, "A study of HMM-based bandwidth extension of speech signals," *Signal Processing*, vol. 89, no. 10, pp. 2036–2044, 2009.

[10] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, 2003.

[11] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep Learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *Signal Processing Magazine, IEEE*, vol. 32, no. 3, pp. 35–52, May 2015.

[12] Y. Gu and Z. Ling, "Restoring high frequency spectral envelopes using neural networks for speech bandwidth extension," in *Neural Networks (IJCNN), 2015 International Joint Conference on*, July 2015, pp. 1–8.

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

[14] Y. Wang, S. Zhao, W. Liu, M. Li, and J. Kuang, "Speech bandwidth expansion based on deep neural networks," in *INTERSPEECH 2015*, September 2015, pp. 2593–2597.

[15] B. Liu, J. Tao, Z. Wen, Y. Li, and D. Bukhari, "A novel method of artificial bandwidth extension using deep architecture," in *INTERSPEECH 2015*, September 2015, pp. 2598–2602.

[16] Y. Wang, S. Zhao, J. Li, and J. Kuang, "Speech bandwidth extension using recurrent temporal restricted boltzmann machines," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1877–1881, Dec 2016.

[17] Y. Gu, Z. Ling, and L. Dai, "Speech bandwidth extension using bottleneck features and deep recurrent neural networks," in *INTERSPEECH 2016*, September 2016, pp. 297–301.

[18] B. Liu and J. Tao, "A novel research to artificial bandwidth extension based on deep BLSTM recurrent neural networks and exemplar-based sparse representation," in *INTERSPEECH 2016*, September 2016, pp. 3778–3782.

[19] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[20] D. Palaz, R. Collobert, and M. Magimai-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *INTERSPEECH 2013*, September 2013, pp. 1766–1770.

[21] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, "Convolutional neural networks for acoustic modeling of raw time signal in LVCSR," in *INTERSPEECH 2015*, September 2015, pp. 26–30.

[22] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs," in *INTERSPEECH 2016*, September 2016, pp. 3434–3438.

[23] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[24] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. v. d. Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," *arXiv preprint arXiv:1610.10099*, 2016.

[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *ICLR*, 2015.

[26] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[27] J. B. Allen and L. R. Rabiner, "A unified approach to short-time fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[29] R. G. ITU-T and I. Switzerland, "711. pulse code modulation (PCM) of voice frequencies," *International Telecommunication Union, Geneva, Switzerland*, 1988.

[30] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.

[31] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "TensorFlow: large-scale machine learning on heterogeneous systems, software available from tensorflow. org, 2015," *Available: http://tensorflow. org*.