



DNN-Based Speech Bandwidth Expansion and Its Application to Adding High-Frequency Missing Features for Automatic Speech Recognition of Narrowband Speech

Kehuang Li¹, Zhen Huang¹, Yong Xu^{2*}, and Chin-Hui Lee¹

¹Georgia Institute of Technology, Atlanta, GA 30332-0250, USA

²University of Science and Technology of China, Hefei, Anhui, P. R. China

{kehle, zhuang41}@gatech.edu, xuyong62@mail.ustc.edu.cn, chl@ece.gatech.edu

Abstract

We propose a number of enhancement techniques to improve speech quality in bandwidth expansion (BWE) from narrowband to wideband speech, addressing three issues, which could be critical in real-world applications, namely: (1) discontinuity between narrowband spectrum and the estimated high frequency spectrum, (2) energy mismatch between testing and training utterances, and (3) expanding bandwidth of out-of-domain speech signals. With an inherent prediction of missing high frequency features in bandwidth-expanded speech we also explore the feasibility of adding these estimated features to those extracted from narrowband speech in order to improve the system performance for automatic speech recognition (ASR) of narrowband speech. Leveraging upon a recently-proposed deep neural network based speech BWE system intended for hearing quality enhancement these techniques not only improve over the traditionally-adopted objective and subjective measures but also reduce the word error rate (WER) from 8.67% when recognizing narrowband speech to 8.26% when recognizing bandwidth-expanded speech, and almost approaching the WER of 8.12% when recognizing wideband speech in the 20,000-word open-vocabulary Wall Street Journal ASR task.

Index Terms: deep neural network, speech bandwidth expansion, automatic speech recognition

1. Introduction

Speech bandwidth expansion (BWE) from narrowband to wideband speech has been studied for decades [1, 2, 3, 4] for the purpose of enhancing the listening quality of narrowband speech, such as that over existing public switching telephone network (PSTN) [5, 6]. Even now the bandwidth for speech transmission is no longer critically limited, we still have plenty of devices and equipments that transmit and receive narrowband speech. For instance most blue-tooth headphones [7] are still working on narrowband speech. Thus expanding speech bandwidth from narrowband (with 4 kHz bandwidth) to wideband (with 8 kHz bandwidth or higher) will still benefit our daily life.

On the other hand, human beings are not the only targets involved in communication that a great amount of research and engineering efforts have been put into human computer interface (HCI) systems [8, 9, 10], that spoken dialog systems [11, 12, 13] are able to understand natural human speech and make response. Thus if bandwidth expansion can improve speech intelligibility for human and computer at the same time,

*This work was done during Yong Xu’s visiting research at Georgia Institute of Technology in 2014-2015

computer aided systems built on narrowband channels can provide better user experience that not only with enhanced speech quality but also with potential improved accuracies in the automatic speech recognition (ASR) [14, 15, 16] component of these dialog systems.

Traditional bandwidth expansion was usually focused on estimating the spectral envelope of the high-frequency band and the excitation of the low-frequency band to recover the high-frequency spectrum [17, 18, 19, 20]. Some recent research also show that estimating the high-frequency spectrum directly is feasible [21, 22]. Our previous work [23] shows that DNN based BWE can work pretty well in this case. However, there is a problem called spectrum discontinuity. In [24], a similar problem was resolved by pivoting the end point of narrowband spectrum and smoothing high-frequency components to fit it.

We propose to jointly estimate the entire wideband spectrum instead of just predicting the high-frequency spectrum, based on the observation that DNN inherently produces a smooth output in most DNN based regression tasks. It is found, contradictory to conventional thinking, that such a DNN framework can remove the transition discontinuity between the narrowband and the high-frequency spectra. It can also reduce the differences between the narrowband spectrum and the estimated low-frequency spectrum of the predicted wideband signal. Another critical issue in BWE is the energy mismatch between the testing and training utterances. A speech activity detection (VAD [25, 26]) and a speech energy adjustment step can help to a certain degree. Leveraging on the property that the low-frequency elements of the cepstral coefficients can characterize the average energy of a given utterance [27] we also propose adding them to DNN training. As a result we find a DNN trained in this way can better handle the problem of the test utterance being out-of-domain, in which the test utterances used for BWE test could acoustically be very different from those used in DNN based BWE training. Moreover, to expand the bandwidth of out-of-domain speech, the channel mismatch issue still has to be addressed. In our experiment, we found that utterance normalization can reduce some channel mismatch modelled by a bias vector in the log spectrum domain.

These three proposed improvement techniques over our recently proposed DNN based BWE system [23] by predicting the overall wideband spectrum in DNN training, adding cepstral features to DNN training, and utterance normalization are shown in our experimental results to be capable of addressing the three critical issues in enhancing human listening quality, namely: (1) discontinuity between narrowband spectrum and the estimated high frequency spectrum, (2) energy mismatch

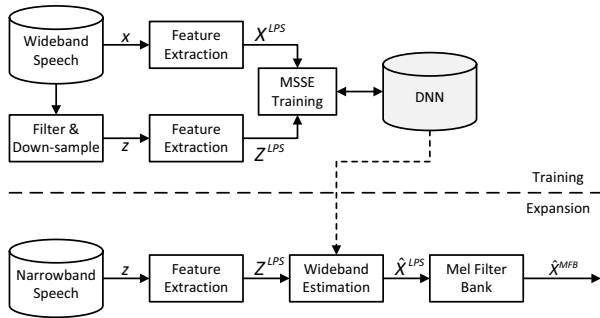


Figure 1: A block diagram of DNN-BWE module.

between testing and training utterances, and (3) ability to expand bandwidth of out-of-domain speech signals. The quality of the enhanced BWE speech signals also demonstrates excellent performances in both subjective and objective tests.

Furthermore, we also explore the feasibility of adding to those already in narrowband speech the speech features embedded in the additional high-frequency spectrum for the purpose of improving the speech recognition system accuracy. Our preliminary experiments on ASR of the 20000-word open vocabulary Wall Street Journal task confirms that wideband speech, with a word error rate (WER) of 8.12%, can always outperform narrowband speech, with a WER of 8.67%, if the systems are based on the same architecture and training strategies. Nonetheless using the same design on the band-expanded speech gives a WER of 8.26% which is close to the performance of wideband speech. This encouraging result will inspire our future work of exploring new algorithms to estimate additional speech features in the missing high frequency spectra in order to take advantage of many existing narrowband speech corpora and even combine them with existing wideband speech databases to further improve the ASR performance.

2. DNN Based Bandwidth Expansion

2.1. Feature Extraction

In our recent work [23], a DNN-based BWE system was proposed. Although good objective and subjective performances have been reported, predicting high half band parameters often led to some discontinuity problem from the transition points between low-frequency (0~4 kHz) to high-frequency (4~8 kHz) spectra when using log-power spectrum (LPS) as features. To reduce the problem, one possible remedy could be smoothing the speech parameters at these transitions by compensating the energy gap. Our recent experiments show that, a DNN-based BWE system can predict the whole wideband spectrum instead of just the missing high-frequency band. By doing so, it will lead to a slight mismatch between the predicted low-frequency band and the original narrowband, which is usually hardly noticed in the spectrogram but can be exposed when objective measures are used.

For BWE, LPS of the narrowband signal, Z^{LPS} , and that of the wideband signal, X^{LPS} , are used as input and output features of the DNN, and zero-mean unit-variance normalization [28] was performed on the features before they are fed into DNN [23]. Let μ_n and μ_w be the mean vectors of the LPS features of the training data, and Σ_n and Σ_w their corresponding variances along each feature dimension. Then the input feature is $(Z^{LPS} - \mu_n) \cdot \Sigma_n^{-1}$, and the output of DNNs, Y , and the

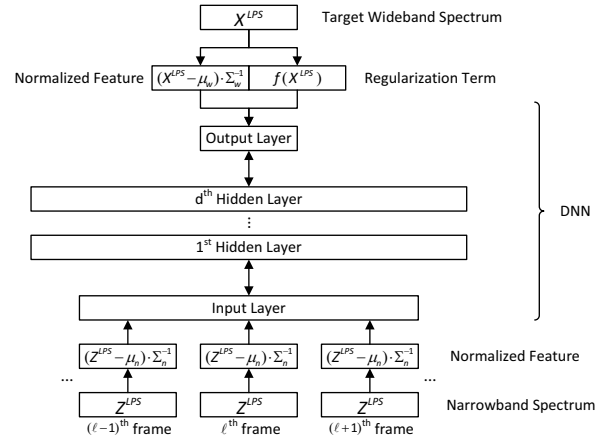


Figure 2: DNN-BWE architecture and training.

estimated wideband LPS, $X^{\hat{LPS}}$, follow the following relation:

$$\hat{X}^{LPS} = Y \cdot \Sigma_w + \mu_w. \quad (1)$$

Some recent research shows that Mel-filter bank [29] features deliver a good performances in DNN based ASR systems. Furthermore we can easily convert LPS to log Mel-filter bank features as follows,

$$X^{MFB} = \ln \left[\exp \left(X^{LPS} \right) \times F \right], \quad (2)$$

where X^{LPS} is the N -dimension wideband LPS feature vector, while F is an $N \times K$ matrix of Mel-filter banks with K filter bins. For the narrowband signals, and the corresponding Mel-filter feature, Z^{MFB} , the filter bank matrix used is different from the one for the wideband signal, since the frequency range and the number of bins are different, and usually the narrowband signal will have less bins. The number of filter bins is decided by making sure the narrowband signal will have most similar filter bins as those of the wideband signal.

2.2. DNN Training

We used the Kaldi toolkit [28] to train DNNs. Unsupervised pre-training of restricted Boltzmann machine (RBM) was first performed [30]. Then, in discriminative fine tuning, a minimum sum of squared error (MSSE [31]) criterion was used in an attempt to minimize the Euclidean distance between the predicted wideband features and the true wideband features of the desired wideband signal. Let $[Y; R]$ be the output of DNN, where R is some extra output vector for regularization purpose that will be truncated from the final output, Y , the objective function is

$$\min \quad \frac{1}{2} \left\| \left(X^{LPS} - \mu_w \right) \Sigma_w^{-1} - Y \right\|_2^2 + \frac{\rho}{2} \left\| f \left(X^{LPS} \right) - R \right\|_2^2 + \frac{\gamma}{2} \left(\|Y\|_2^2 + \|R\|_2^2 \right), \quad (3)$$

where ρ is penalty ratio. The third term is L-2 penalty of the overall output vector, but γ was set to 0 in this work. $f(\cdot)$ can be any function and if $f(X^{LPS})$ is a target of other purpose than BWE, it turns out to be multi-task learning [32]. Here we used lower half of the cepstrum in the second term, that is $f(\cdot)$ is doing discrete cosine transform (DCT) and discarding higher half of the parameters. In this way, we can better handling the energy mismatch between the narrowband and wideband spectra,

which could be tough when the input narrowband signal has a large bias from the training data.

3. Speech Recognition on BWE Speech

The acoustic model adopted in this paper for ASR is also feed-forward DNNs [33]. We first pass the narrowband signal’s spectrum through the DNN-based BWE system to get the estimated wideband spectrum, then extract the feature vector (log Mel-filter bank) using the estimated wideband spectrum to feed into the DNN acoustic model for ASR.

In a typical setting of the DNN acoustic model, the hidden layers are usually constructed by sigmoid units, and the output layer is a soft-max layer directly modelling tied context-dependent triphone states, sometimes referred to as senones [34]. The DNN was trained by maximizing the log posterior probability over the training frames. This is equivalent to minimizing the cross-entropy objective function. Let \mathcal{X} be the whole training set, which contains T frames, *i.e.*, $\mathbf{o}^{1:T} \in \mathcal{X}$, then the loss with respect to \mathcal{X} is given by

$$\mathcal{L}^{1:T} = - \sum_{t=1}^T \sum_{j=1}^J \tilde{\mathbf{p}}^t(j) \log p(C_j | \mathbf{o}^t), \quad (4)$$

where $p(C_j | \mathbf{o}^t)$ is the posterior probability of senone j ; $\tilde{\mathbf{p}}^t$ is the target probability of frame t . In real practices of DNN systems, the target probability $\tilde{\mathbf{p}}^t$ is often obtained by a forced alignment with an existing system resulting in only the target entry that is equal to 1. Mini-batch stochastic gradient descent (SGD) [35], with a reasonable size of mini-batches to make all matrices fit into the GPU memory, was used to update all neural parameters during training. Pre-training methods was used for the initialisation of the DNN parameters.

4. Experiments and Result Analysis

4.1. Experimental Setup

We experimented on the Wall Street Journal (WSJ0) corpus [36] with microphone speech sampled at 16 kHz in 16 bits resolution, and the Switch Board (SWB1) corpus [37] at 8 kHz in 16 bits resolution. WSJ0 with 31166 utterances in the training set (with about 50 hours for training and 10 hours for validation) was used to train the BWE model. The window size of STFT was 400 samples with a shift length of 160 samples on the wideband signal, while the narrowband signal took a window size of 200 with a window shift of 80. Hamming window was used in feature extraction. Mel-filter bank for the wideband signal had 29 bins from 0 Hz to 8000 Hz, and the first 22 bins cover from 0 Hz to 4132 Hz. Narrowband features took coefficients of 22 filter bank bins that cover the frequency range of 0 Hz to 4000 Hz. The base learning rate of MSSE training was set to 10^{-5} , and the “newbob” method [38] was applied that halves the learning rate when the decrease of the mean squared error is less than 0.25%, and stops when it’s less than 0.25%.

Table 1: LSD of the WSJ Testing Set.

	LSD (dB)	LSD _H (dB)
HB	6.13	7.73
WB	5.45	7.60
WB+Cep	5.42	7.58

Mini-batch training [39] with a batch size of 32 utterances and momentum rate of 0.9 was adopted.

For the ASR experiments on the 20k-word open vocabulary Wall Street Journal task, the DNN acoustic model was trained using the WSJ0 material (SI-84). The standard adaptation set of WSJ0 (si_et_ad, 8 speakers, 40 sentences per speaker) was used to perform adaptation of the affine transformation added to the speaker-independent DNN. The standard open vocabulary 20,000-word (20k) read NVP Senneheiser microphone (si_et_20, 8 speakers \times 40 sentences) data were used for evaluation. A standard trigram language model was adopted during decoding. The ASR performance is given in terms of the word error rate (WER).

4.2. Objective Evaluation of DNN-Based BWE Model

We adopted the DNN structure settings in [23, 40], with 11 frames at the input, and 3 hidden layers with 2048 hidden nodes per layer. Here 11 frames means that 5 preceding and 5 following frames were concatenated together with the current frame to feed into the input layer of DNNs. The objective measures are log-spectral distortion (LSD) [41] and segmental signal-to-noise ratio (SegSNR) [42] illustrated in [23]. Results are listed in Table 1 and Table 2, where LSD_H is the LSD of high frequency spectrum, “HB” means DNN was trained to predict high frequency parameters only, “WB” means DNN would predict whole wideband parameters, and “WB+Cep” means DNN learnt both LPS and cepstral parameters.

The LSD results on the left column in Table 1 show that when DNN was trained to predict the whole wideband parameters (“WB”), it could actually reduce the overall spectral difference between narrowband spectrum and the low frequency spectrum of the wideband signal (from 6.13 dB obtained with “HB” to 5.45 dB), where the difference could be caused by the channel (simulated as a lowpass filter or bandpass filter) used to generate the narrowband signal from the wideband signal. Furthermore, learning cepstral together with LPS parameters (“WB+Cep”) further reduced the LSD to 5.42 dB. Similar conclusions can be drawn for the LSD_H (high frequency LSD) results on the right column in Table 1.

Moreover the proposed “WB+Cep” DNN training strategy achieved great gains over “WB” (proposed in this study) and “HB” (proposed in [23]) on SegSNR as shown in Table 2 with only results of reconstructed signals given the real phase of the wideband signal (“CP”) and the high-frequency imaged phase from the narrowband signal (“IP” as in [23]). Clearly the “CP” case on the left column was improved by 0.95 dB (from 18.13 to 19.08 dB). Even for the “IP” case on the right column, when the phase was not re-estimated, we still obtained a gain of 0.6 dB (from 13.28 to 13.88 dB).

4.3. Subjective Test on Out-of-Domain Dataset

DNN models are known to deliver a strong learning ability and robustness. This motivated us to test on out-of-domain narrowband speech. The data set we used to train the BWE system was

Table 2: SegSNR of the WSJ Testing set.

	SegSNR_CP (dB)	SegSNR_IP (dB)
HB	16.47	12.78
WB	18.13	13.28
WB+Cep	19.08	13.88

WSJ0 which is relatively clean and stable, and the data set we chose to cross test the system is the SWB set which is relatively tough that it might contain other voices than speech with echo and sometimes multiple speakers. To reduce the channel mismatch, which could be the most significant problem if we use mismatch data sets for training and testing, normalization was performed at the utterance level in this case, that each utterance will have its own μ_n^i and Σ_n^i to be normalized.

We also performed utterance level normalization on the training data, and retrained DNN using the new features but with the same DNN structure. Since the SWB data are real-world narrowband speech and its bandwidth was only 3.4 kHz, we built the BWE system from 3.5 kHz to 7 kHz, and thus the feature dimensions should be modified correspondingly. On the other hand, there is no μ_w nor Σ_w available for the SWB utterances, and without them it is not easy to reconstruct LPS and waveforms afterwards. Here we borrowed μ_w and Σ_w estimated from the WSJ data, and bias b_i and scale s_i were applied on them respectively. Assume μ_{wn} is the low frequency part of μ_w , and Σ_{wn} is the low frequency part of Σ_w , then we have,

$$b_i = \overline{\mu_n^i - \mu_{wn}} \quad (5)$$

$$s_i = \left| \Sigma_n^i \cdot \Sigma_{wn}^{-1} \right|^{\frac{1}{M}}, \quad (6)$$

where i is the index of the utterance, $\overline{\cdot}$ gives the mean of all entries of a vector, $|\cdot|$ is the determinant, and M is the dimension of the narrowband feature vectors. Then the method in [23] with the narrowband phase (“IP”) was used to reconstruct the waveforms for listening tests.

An example SWB utterance is shown in Figure 3. If only global normalization was used (upper right panel), the estimated high-frequency spectrum has lower clearness with less energy, and the elliptical area shows high-frequency missing for several frames, which might come from channel mismatch that DNN failed to find a pattern for them. If utterance level normalization was used (lower left panel), there are more energy in the high-frequency spectrum, but the rectangular area shows noise being expanded to a consonant. We believe it mostly came from the energy mismatch that training data did not have noise (non-speech sound) at the same energy level. The proposed one (lower right panel) shows an excellent performance on all sounds that the spectrogram is more clear and accurate.

We invited male and female students around our lab at Georgia Tech to test their preference to bandwidth-expanded speech. None of the volunteers work on BWE. A subset with 1% of the testing set of SWB was randomly selected, and each volunteer was asked to listen to 10 sequences which were randomly picked among the 1% subset. In the preference test volunteers were given two parallel sequences at one time without other information about the sequences, and were asked to choose their preferred one or “No Preference” (N/P). Results are shown in Table 3, and in most cases bandwidth expanded sequences were chosen as having better quality (almost 90%) and higher clearness than the other two options. Only in the few cases when volunteers prefer narrowband speech, it was described as being with less noise or giving a softer quality.

Table 3: Preference Test on SWB.

Narrowband	BWE	N/P
5.7%	88.6%	5.7%

4.4. ASR on 20000-word Open-Vocabulary WSJ Task

Finally we gave a preliminary ASR test on the 20k-word open vocabulary Wall Street Journal task, in which the DNN acoustic model was trained using the WSJ0 material (SI-84). Original wideband speech, narrowband speech that only has low-frequency spectrum (0~4 kHz) of the wideband speech, and the speech bandwidth-expanded from narrowband speech (“WB” in 4.2) were used to train the DNN based acoustic models of the ASR system separately using the Kaldi toolkit [28]. Then three models were used to decode speech on the test set and the word error rates are listed in Table 4. Original wideband speech achieved 8.12% WER, and narrowband speech got a 8.67% WER, or a 6.8% relative WER degradation, while BWE speech obtained a 8.26% WER, or a 4.7% relative improvement on narrowband speech and only a 1.7% WER degradation when compared with that obtained with original wideband speech.

Table 4: WER obtained from the wideband, bandwidth-expanded and narrowband speech training data in 20k-word open vocabulary Wall Street Journal ASR task.

	Wideband	BWE	Narrowband
WER	8.12%	8.26%	8.67%

5. Conclusion and Future Work

In this paper, a DNN based BWE framework was explored to improve the performance of BWE in both hearing quality and speech recognition. Three new techniques were explored to resolve the spectrum discontinuity issue mentioned in our previous work, and to improve the system performance when there is energy mismatch and channel mismatch. Experiment results showed that the objective measures achieved significant improvements on in-domain test utterances, and that subjective listening tests on out-of-domain narrowband utterances further confirm the improved system performance in real-world application. Furthermore, the preliminary ASR experiments show that the BWE techniques can improve the recognition performance of narrowband speech. Future work will focus on migrating the BWE-enhanced ASR framework to handle out-of-domain real-world data sets through additional channel matching and normalization.

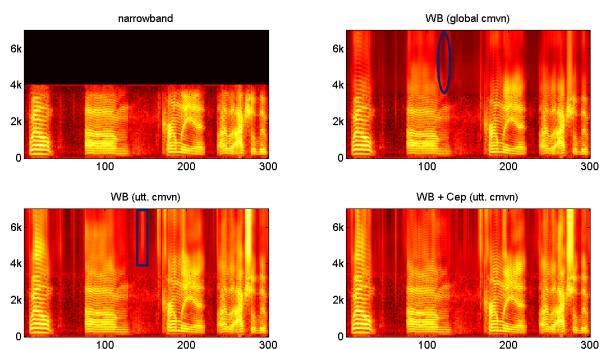


Figure 3: Spectrograms of an example SWB utterance: upper left – original narrowband speech, upper right – bBWE speech with global normalization, lower left – BWE speech using utterance normalization, and lower right – BWE using utterance normalization and cepstral multi-task learning.

6. References

- [1] J. Epps and W. H. Holmes, "A new technique for wideband enhancement of coded narrowband speech," in *Proc. IEEE Workshop on Speech Coding*, 1999, pp. 174–176.
- [2] S. Vaseghi, E. Zavarzani, and Q. Yan, "Speech bandwidth extension: extrapolations of spectral envelope and harmonicity quality of excitation," in *Proc. ICASSP*, vol. 3, 2006.
- [3] J. Han, G. J. Mysore, and B. Pardo, "Language informed bandwidth expansion," in *Proc. IEEE Workshop on Machine Learning for Signal Processing*, 2012, pp. 1–6.
- [4] H. Seo, H.-G. Kang, and F. Soong, "A maximum a posteriori-based reconstruction approach to speech bandwidth expansion in noise," in *Proc. ICASSP*, 2014, pp. 6087–6091.
- [5] L. A. Litteral, J. B. Gold, D. C. Klika Jr, D. B. Konkle, C. D. Coddington, J. M. McHenry, and A. A. Richard III, "PSTN architecture for video-on-demand services," 1993, US Patent 5,247,347.
- [6] C. Marvin, *When old technologies were new*. Oxford University Press, 1997.
- [7] J. C. Haartsen and E. Radio, "The bluetooth radio system," in *IEEE Personal Communications*, 2000.
- [8] B. Laurel and S. J. Mountford, *The art of human-computer interface design*. Addison-Wesley Longman Publishing Co., Inc., 1990.
- [9] B. Shneiderman, *Designing the user interface: strategies for effective human-computer interaction*. Addison-Wesley Reading, MA, 1992, vol. 2.
- [10] R. Sharma, V. Pavlovic, and T. S. Huang, "Toward multimodal human-computer interface," *Proc. of the IEEE*, vol. 86, no. 5, pp. 853–869, 1998.
- [11] V. Zue, S. Seneff, J. R. Glass, J. Polifroni, C. Pao, T. J. Hazen, and L. Hetherington, "JUPITER: a telephone-based conversational interface for weather information," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 1, pp. 85–96, 2000.
- [12] J. D. Williams and S. Young, "Partially observable Markov decision processes for spoken dialog systems," *Computer Speech & Language*, vol. 21, no. 2, pp. 393–422, 2007.
- [13] R. Pieraccini and J. Huerta, "Where do we go from here? research and commercial spoken dialog systems," in *6th SIGdial Workshop on Discourse and Dialogue*, 2005.
- [14] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *The Bell System Technical Journal*, vol. 62, no. 4, pp. 1035–1074, 1983.
- [15] C.-H. Lee, F. K. Soong, and K. K. Paliwal, *Automatic speech and speaker recognition: advanced topics*. Springer Science & Business Media, 1996, vol. 355.
- [16] C. Weng, D. Yu, S. Watanabe, and B.-H. F. Juang, "Recurrent deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2014, pp. 5532–5536.
- [17] B. Iser and G. Schmidt, "Bandwidth extension of telephony speech," in *Speech and Audio Processing in Adverse Environments*. Springer, 2008, pp. 135–184.
- [18] Y. Nakatoh, M. Tsushima, and T. Norimatsu, "Generation of broadband speech from narrowband speech based on linear mapping," *Electronics and Communications in Japan (Part II: Electronics)*, vol. 85, no. 8, pp. 44–53, 2002.
- [19] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. ICASSP*, vol. 3, 2000, pp. 1843–1846.
- [20] 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.
- [21] K. Kalgaonkar and M. A. Clements, "Sparse probabilistic state mapping and its application to speech bandwidth expansion," in *Proc. ICASSP*, 2009, pp. 4005–4008.
- [22] H. Seo, H.-G. Kang, and F. Soong, "A maximum a posteriori-based reconstruction approach to speech bandwidth expansion in noise," in *Proc. ICASSP*, 2014, pp. 6087–6091.
- [23] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. ICASSP*, 2015.
- [24] B. Iser and G. Schmidt, "Neural networks versus codebooks in an application for bandwidth extension of speech signals," in *Proc. INTERSPEECH*, 2003, pp. 565–568.
- [25] J. Ramirez, J. M. Górriz, and J. C. Segura, *Voice activity detection: Fundamentals and speech recognition system robustness*. INTECH Open Access Publisher, 2007.
- [26] S. Morita, M. Unoki, X. Lu, and M. Akagi, "Robust voice activity detection based on concept of modulation transfer function in noisy reverberant environments," in *Chinese Spoken Language Processing (ISCSLP), 9th International Symposium on*, 2014, pp. 560–564.
- [27] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1, pp. 133–147, 1998.
- [28] 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 *Proc. ASRU*, 2011, pp. 1–4.
- [29] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *Proc. ICASSP*, vol. 1, 2001, pp. 73–76.
- [30] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for Boltzmann machines," *Cognitive Science*, vol. 9, no. 1, pp. 147–169, 1985.
- [31] D. M. Allen, "Mean square error of prediction as a criterion for selecting variables," *Technometrics*, vol. 13, no. 3, pp. 469–475, 1971.
- [32] R. Caruna, "Multitask learning: A knowledge-based source of inductive bias," in *Proc. ICML*, 1993, pp. 41–48.
- [33] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 1, pp. 30–42, 2012.
- [34] M.-Y. Hwang and X. Huang, "Shared-distribution hidden Markov models for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 4, pp. 414–420, 1993.
- [35] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao, "Optimal distributed online prediction," in *Proc. ICML*, 2011, pp. 713–720.
- [36] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *HLT '91 Proceedings of the workshop on Speech and Natural Language*, 1992, pp. 357–362.
- [37] J. J. Godfrey and E. Holliman, "Switchboard-1 release 2," *Linguistic Data Consortium, Philadelphia*, 1997.
- [38] ICSI QuickNet toolbox. Newbob approach is implemented in the toolbox. [Online]. Available: <http://www1.icsi.berkeley.edu/Speech/qn.html>
- [39] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," Dept. Comput. Sci., Univ. Toronto, Tech. Rep. UTML TR 2010-003, 2010.
- [40] Y. Xu, J. Du, L. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, pp. 65–68, 2014.
- [41] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 873–902.
- [42] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall Englewood Cliffs, NJ, 1988.