A DEEP NEURAL NETWORK APPROACH TO SPEECH BANDWIDTH EXPANSION

Kehuang Li *Chin-Hui Lee*

School of ECE, Georgia Institute of Technology, Atlanta, GA 30332-0250, USA
kehle@gatech.edu, chl@ece.gatech.edu

ABSTRACT

We propose a deep neural network (DNN) approach to speech bandwidth expansion (BWE) by estimating the spectral mapping function from narrowband (4 kHz in bandwidth) to wideband (8 kHz in bandwidth). Log-spectrum power is used as the input and output features to perform the required non-linear transformation, and DNNs are trained to realize this high-dimensional mapping function. When evaluating the proposed approach on a large-scale 10-hour test set, we found that the DNN-expanded speech signals give excellent objective quality measures in terms of segmental signal-to-noise ratio and log-spectral distortion when compared with conventional BWE based on Gaussian mixture models (GMMs). Subjective listening tests also give a 69% preference score for DNN-expanded speech over 31% for GMM when the phase information is assumed known. For tests in real operation when the phase information is imaged from the given narrowband signal the preference comparison goes up to 84% versus 16%. A correct phase recovery can further increase the BWE performance for the proposed DNN method.

Index Terms— Deep neural network, speech bandwidth expansion, spectrum mapping, phase estimation

1. INTRODUCTION

Expanding speech bandwidth from narrowband (with 4 kHz bandwidth) to wideband (with 8 kHz bandwidth) has been studied for decades as the bandwidth was an expensive resource in the early years. Even now the bandwidth for speech transmission is no longer tensely limited, we still face low bandwidth restriction in the existing public switching telephone network (PSTN) system. To enhance the listening quality of speech over PSTN, efforts have been made to artificially extend the bandwidth.

Many early studies on bandwidth expansion (BWE) focused on estimating the spectral envelope of the high-frequency band, and using the excitation generated from the low-frequency band to recover the high-frequency spectrum [1]. A few techniques, such as linear mapping [2], piecewise linear mapping [3, 4], codebook mapping [5, 6], neural networks [7, 8], Gaussian mixture model [9, 10], and hidden Markov model [11, 12] and non-negative hidden Markov model [13], have been explored. Linear predictive coefficients (LPCs) or line spectral frequencies (LSFs) [14, 15] are widely used to represent the spectral envelope, while the excitation can be found by inverse filtering the signal with LPCs, modulation techniques, non-linear processing, and the application of function generators [1].

In contrast to envelope estimation methods, direct estimation of the missing high-frequency spectrum was not extensively studied because the dimensions of both original low-frequency spectrum and target high-frequency spectrum spaces to establish a mapping function are really high. However, there are still some studies, such as folded spectrum adjusting [8] and sparse probabilistic state mapping [16]. The former one folds the narrowband spectrum and adjusts the level of the wideband spectrum, attempting to estimate the spectral envelope in a different way. The latter one assumes that the transmission matrix of the mapping is sparse, which is usually inaccurate. However, these techniques show that a direct spectrum estimation of the missing band can have some benefits and is worth further study.

In summary, we propose to use DNN for spectral mapping to estimate the missing high-frequency spectrum. Experiments on a large scale 10-hour test set show that the proposed DNN framework demonstrates better objective measures in terms of segmental signal-to-noise ratio (SegSNR) [17] and log-spectral distortion (LSD) [18] when compared to conventional GMM-based mapping techniques. Subjective preference listening tests also give a 69% score over 31% for GMM-expanded speech when the phase information is assumed known. For real operation tests when the phase information is imaged from the given narrowband signal the preference comparison goes up to 84% versus 16%. A correct phase recovery can further increase the BWE performance for the proposed DNN approach.

2. DNN BASED SPEECH BANDWIDTH EXPANSION

2.1. Feature Extraction

A block diagram of the proposed DNN-based BWE system is shown in Figure 1. Given a wideband speech signal x, we windowed it into overlapping frames, and performed a short-time Fourier transform (STFT) [19] on the windowed frame

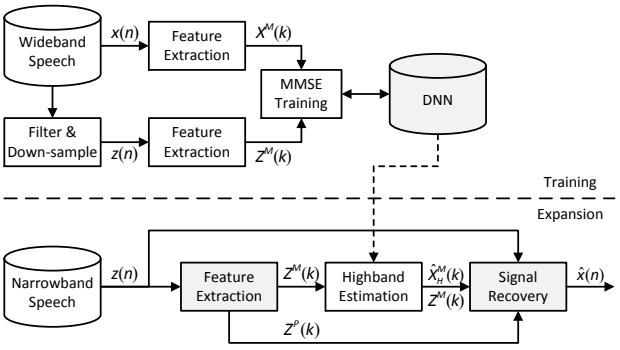


Fig. 1. A block diagram of the proposed DNN-BWE system.

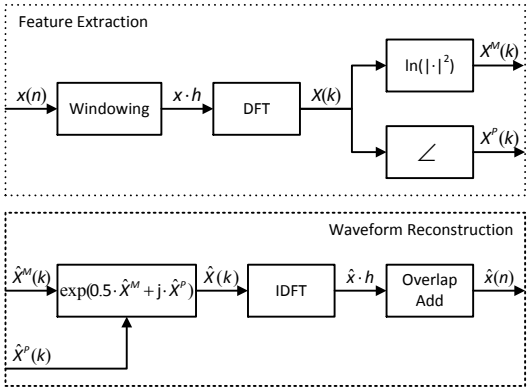


Fig. 2. A flowchart of feature extraction and wave reconstruction.

as follows,

$$X(\ell,k)=\sum_{n=0}^{N-1}x(\ell\times\Delta+n)h(n)e^{-j2\pi nk/N}, \quad (1)$$

where ℓ is the frame index, $k=0,...,L-1$ is the discrete frequency index, Δ is the window shift, N is the window length, and $h(\cdot)$ denotes the window function, which is a Hamming window here. We will omit ℓ in the rest of the article as we focus on features of each frame. Log-spectral power magnitude was then extracted [20],

$$X^M(k)=\ln|X(k)|^2. \quad (2)$$

Since x is a real signal, X is conjugate symmetric and is uniquely determined by only $N/2+1$ points. Thus we use $X^M(k)$ with $k=0,...,N/2$ as features. For the wideband signal, X^M was further separated into a low-frequency spectrum, $X_L^M=[X^M(0),...,X^M(N/4)]$, and a high-frequency spectrum, $X_H^M=[X^M(N/4+1),...,X^M(N/2)]$, where X_H^M is to be recovered by DNN based on the narrowband (low-frequency) spectrum.

Besides the magnitude of the Fourier coefficients, the phase information was extracted as follows,

$$X^P(k)=\angle X(k). \quad (3)$$

As for the wideband signal, X^P was separated to X_L^P and X_H^P in the same way as its corresponding magnitude X^M.

A narrowband signal z was generated by filtering and down-sampling the wideband signal x, and Z^M and Z^P are its corresponding log-spectral magnitude and phase.

2.2. DNN Training

As shown in Figure 3, the input of the DNN is the log-spectrum of the narrowband signal and the output is the high-frequency log-spectrum of the wideband signal. To ensure the proper working of DNNs, each dimension of DNNs' input and output was normalized among all training samples to ensure it is of zero mean and unit variance. Thus in the application stage of bandwidth expansion, the same normalizing step was executed on input feature vectors, and a reverse step on the output is necessary.

We used the Kaldi toolkit [21] to train DNNs. Unsupervised pre-training of restricted Boltzmann machine (RBM) was first performed [22]. Then, in discriminative fine tuning, the minimum mean square error (MMSE) criterion was used in an attempt to minimize the Euclidean distance between the predicted high-frequency log-spectrum and the true high-frequency log-spectrum of the desired wideband signal. Let Y be the output of DNN, and the objective function of MMSE is

$$\min\ \frac{1}{2}\left\|\left(X_H^M-\mu_H\right)\Sigma_H^{-1}-Y\right\|_2^2, \quad (4)$$

where μ_H and Σ_H^{-1} are the mean vector and the diagonal inverse covariance matrix of all high-frequency log-spectrum of training data.

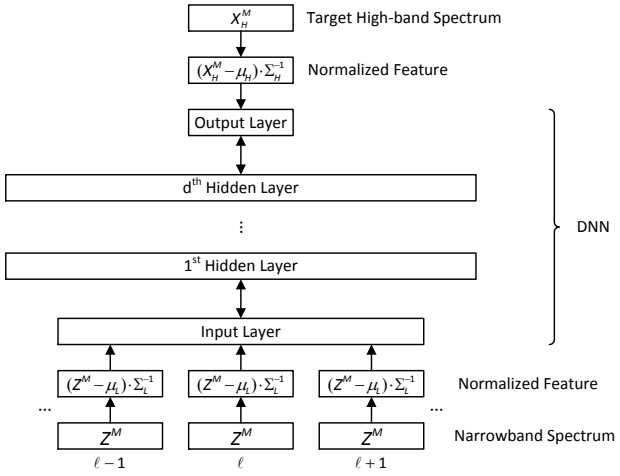


Fig. 3. DNN architecture and training.

2.3. Waveform Reconstruction

Even if it was possible to obtain the exact magnitude of the wideband spectrum, the phase information was lost in the previous steps. Based on DNN's output, we have an estimation of the high-frequency spectrum $\hat{X}_H^M = (Y + \mu_H)\,\Sigma_H$, and $\hat{X}^M = [Z^M + 2\ln 2, \hat{X}_H^M]$, an estimation of expanded wideband spectrum, where $2\ln 2$ compensates the energy loss due to only half of the points of wideband signal is used to calculate the narrowband spectrum. The narrowband spectrum is not modified in order to prevent quality degradation [7]. As for the phase, we have an estimation of the low-frequency phase $\hat{X}_L^P = Z^P$ and the high-frequency phase is unknown. Imaged phase is a simple estimation that $\hat{X}^P = [Z^P, -\text{flip}(Z^P)]$, where $\text{flip}(Z^P)$, or abbreviated as Z_f^P, is defined as $Z_f^P(k) = Z^P(N/4-1-k)$ for $k = 0, 1, ..., N/4-1$. The inverse discrete Fourier transform (IDFT) was then performed on

$$\hat{X}(k) = \exp\left\{\frac{1}{2}\hat{X}^M(k) + j\hat{X}^P(k)\right\}, \tag{5}$$

an inverse step of (2) and (3), and overlap add given in [23] with the same Hamming window for feature extraction was used to reconstruct the signal $\hat{x}$.

3. EXPERIMENTS AND RESULTS

3.1. Experimental Setup

We experimented on the Wall Street Journal (WSJ0) corpus [24] with microphone speech sampled at 16 kHz in 16 bits resolution. A direct comparison with other techniques is not easy. Instead we conducted a large-scale test on WSJ0 with 31166 utterances in the training set (with about 50 hours for training and 10 hours for validation), and 4137 utterances for testing (about 10 hours). The window size of STFT is 512 samples with a shift length of 256 samples on the wideband signal, while the narrowband signal has a window size of 256 with a window shift of 128. The base learning rate of MMSE training was set to 10^{-5}, and the "newbob" method [25] was applied that halves the learning rate when the decrease of the mean squared error is less than 0.1, and stops when it's less than 0.01. Mini-batch training [26] with a batch size of 32 utterances was adopted. As a comparison, a full covariance joint GMM with 2045 mixtures was built and used to perform the same regression function as DNNs.

3.1.1. Objective Quality Measures

The objective quality measures used in our experiments were segmental SNR (SegSNR) [17] and log-spectral distortion (LSD) [18] defined as follows:

$$\text{SegSNR} = \frac{1}{L}\sum_{\ell=0}^{L-1}\left\{10\lg\frac{\sum_{n=0}^{N-1}[x(\ell,n)]^2}{\sum_{n=0}^{N-1}[x(\ell,n)-\hat{x}(\ell,n)]^2}\right\}, \tag{6}$$

where ℓ indicates the ℓ-th frame, and L denotes the number of frames in the utterance.

$$\text{LSD} = \frac{1}{L}\sum_{\ell=0}^{L-1}\left\{\frac{1}{\frac{N}{2}+1}\sum_{k=0}^{\frac{N}{2}}\left[X^M(\ell,k)-\hat{X}^M(\ell,k)\right]^2\right\}^{\frac{1}{2}}, \tag{7}$$

To measure the performance on estimating the high-band spectrum, we also introduce LSD_H that only sums up the distortion in the high half-band with discrete frequency indices, $k = N/4+1, ..., N/2$.

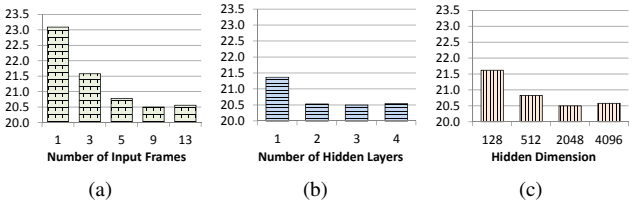
3.1.2. Subjective Test

Besides the aforementioned objective measures, a subjective listening test was conducted as well. Ten volunteers were asked to listen to 10 random pairs of test utterances and their preferences were recorded and summed up to indicate the overall preference.

3.2. Results and Discussion

3.2.1. Structure of DNN

The size and shape of DNNs will affect the performance of the neural networks. For simplicity, we focused on DNNs with the same width for all hidden layers. We adopted the structure settings in [27]. As shown in Figure 4, DNN with 9 frames, 3 hidden layers and 2048 hidden nodes per layer is a locally optimal parameter setting in our experiments on the WSJ0 dataset. Here 9 frames means 4 previous and 4 following frames were concatenated together with the current frame to feed into the input layer of DNNs. The performance was seen to be not sensitive to small parameter differences.



(a) (b) (c)

Fig. 4. MSE of different DNNs. Default parameters are 9 frames, 3 layers and 2048 hidden nodes in each layer, and only one of the three parameters is varied in each comparison.

Table 1. Objective Measure on Reconstructed Signals

		SegSNR (dB)	LSD (dB)	LSD_H (dB)
GMM	CP	15.42	6.34	8.28
	IP	12.12	7.29	9.72
DNN	CP	16.47	5.32	6.69
	IP	12.78	6.44	8.44

3.2.2. *Objective Performance*

Table 1 lists the results of SegSNR and LSD of the reconstructed signals of different methods and phase used. The first row of each method is "CP" which indicates we used the "cheated phase", i.e., we used the higher half-band phase of the original wideband signal not available at the input narrowband signal. The second row of each method is "IP" which indicates the use of imaged phases, i.e., we flipped the phase of the input narrowband signal to the upper half-band and added a minus sign to them. Contrary to conventional thinking, if incorrect phase is used in reconstruction, the SegSNR of the reconstructed signal will be greatly degraded (from 15.42 to 12.12 dB for GMM and from 16.47 to 12.78 dB for DNN). The LSD of the cheated phase is always more than 1 dB better than that of the imaged phase. As for LSD of the high-band, it is always more serious than that for the whole band at about 1.3 to 2 dB degradation. Moreover, DNN outperformed GMM in both "CP" and "IP" cases and on all three measures.

Figure 5 gives an example of one female utterance and one male test utterance. One problem of directly predicting high-frequency spectrum is that there will be a discontinuity between low-frequency and high-frequency spectrum.

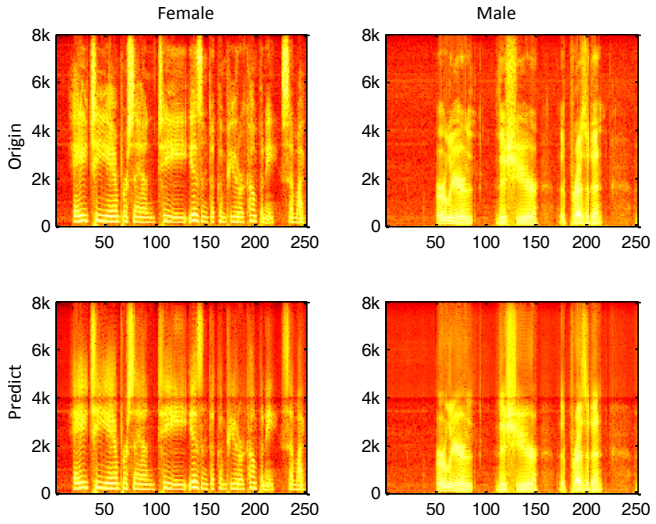


Fig. 5. Spectrogram of one female test utterance and one male test utterance, top row: original, bottom row: reconstructed signal, left column: female, and right column: male.

Table 2. Preference Test on Reconstructed Signals

	CP	IP
GMM	31%	16%
DNN	69%	84%

3.2.3. *Subjective Performance*

When competing with narrowband signal, both GMM and DNN get 100 percent of preference. And Table 2 shows the competition result between GMM and DNN. They give a 69% score over 31% for GMM-expanded speech when the phase information is assume known. For real operation tests when the phase information is imaged from the given narrowband signal the preference comparison goes up to 84% versus 16%. Good estimated phase information can further increase the BWE performance for the proposed DNN approach.

3.2.4. *Comparison of Computation Complexity*

A workstation with 32 2.93 GHz CPU cores and one GTX480 graphic card was used in our experiments. Table 3 shows the computational time of DNN and GMM in training and expansion stages. The test data size itself was about 600 minutes, that is the expansion stage can be real-time. However, the lag time of DNNs depends on the parameter settings. Using the experimental setting mentioned above the lag time was 96 ms ((4 frame shift $\times$ 128 point/shift + 1 current frame $\times$ 256 point/frame) $\div$ 8 kHz).

Table 3. Time Consumption of Training and Testing

	Train	Test
DNN	1501 min	93 min
GMM	358 min	367 min

4. CONCLUSION AND FUTURE WORK

In this paper, a deep neural network based framework of speech bandwidth expansion is proposed. Taking advantage of the deep learning ability, the DNN is shown to be able to map the magnitude spectrum of the input narrowband signal to that of the high-band of the wideband signal. Experimental results, on a large-scale 10-hour test set, show that the proposed DNN framework can effectively estimate the high-frequency spectrum and achieve a higher segmental SNR and lower log-spectral distortion when compared to a GMM based-BWE approach. A subjective test also confirms that our proposed framework demonstrates higher listening preference than the competing GMM-based systems. For further work, we intend to address the spectrum discontinuity issue mentioned in Figure 5. Furthermore with a correct phase recovery we observe that the system performance can further be improved which will be studied in another upcoming paper.

5. REFERENCES

[1] B. Iser and G. Schmidt, "Bandwidth extension of telephony speech," in *Speech and Audio Processing in Adverse Environments*, pp. 135–184. Springer, 2008.

[2] 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.

[3] Y. Nakatoh, M. Tsushima, and T. Norimatsu, "Generation of broadband speech from narrowband speech using piecewise linear mapping," in *Proc. EUROSPEECH*, 1997, pp. 1643–1646.

[4] 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.

[5] U. Kornagel, "Spectral widening of telephone speech using an extended classification approach," in *Proc. EUSIPCO*, 2002, vol. 2, pp. 339–342.

[6] S. Vaseghi, E. Zavarehei, and Q. Yan, "Speech bandwidth extension: extrapolations of spectral envelop and harmonicity quality of excitation," in *Proc. ICASSP*, 2006, vol. 3.

[7] 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.

[8] J. Kontio, L. Laaksonen, and P. Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 15, no. 3, pp. 873–881, 2007.

[9] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. ICASSP*, 2000, vol. 3, pp. 1843–1846.

[10] H. Seo, H.-G. Kang, and F. Soong, "A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise," in *Proc. ICASSP*, 2014, pp. 6087–6091.

[11] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden Markov model," in *Proc. ICASSP*, 2003, vol. 1, pp. I–680.

[12] 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.

[13] 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.

[14] W. P. LeBlanc, B. Bhattacharya, S. A. Mahmoud, and V. Cuperman, "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 4, pp. 373–385, 1993.

[15] F. K. Soong and B.-H. Juang, "Optimal quantization of LSP parameters," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 1, pp. 15–24, 1993.

[16] K. Kalgaonkar and M. A. Clements, "Sparse probabilistic state mapping and its application to speech bandwidth expansion," in *Proc. ICASSP*, 2009, pp. 4005–4008.

[17] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*, Prentice Hall Englewood Cliffs, NJ, 1988.

[18] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, pp. 873–902. Springer, 2008.

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

[20] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions.," in *Proc. INTERSPEECH*, 2008, pp. 569–572.

[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 *Proc. ASRU*, 2011, pp. 1–4.

[22] 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.

[23] D. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, 1984.

[24] 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.

[25] ICSI QuickNet toolbox. Newbob approach is implemented in the toolbox. [Online]. Available: http://www1.icsi.berkeley.edu/Speech/qn.html

[26] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," Tech. rep. utml tr 2010–003, Dept. Comput. Sci., Univ. Toronto, 2010.

[27] 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.