Speech bandwidth expansion based on Deep Neural Networks

Yingxue Wang[1,2], *Shenghui Zhao*[1], *Wenbo Liu*[3,4], *Ming Li*[3,5],*Jingming Kuang* [1]

[1]School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
[2]School of Computer Science, Carnegie Mellon University
[3]SYSU-CMU Joint Inst. of Eng., Sun Yat-Sen University
[4]Department of ECE, Carnegie Mellon University
[5]SYSU-CMU Shunde International Joint Research Institute

yxwang.bit@gmail.com, shzhao@bit.edu.cn

10.21437/Interspeech.2015-558

Abstract

This paper proposes a new speech bandwidth expansion method, which uses Deep Neural Networks (DNNs) to build high-order eigenspaces between the low frequency components and the high frequency components of the speech signal. A four-layer DNN is trained layer-by-layer from a cascade of Neural Networks (NNs) and two Gaussian-Bernoulli Restricted Boltzmann Machines (GBRBMs). The GBRBMs are adopted to model the distribution of spectral envelopes of the low frequency and the high frequency respectively. The NNs are used to model the joint distribution of hidden variables extracted from the two GBRBMs. The proposed method takes advantage of the strong modeling ability of GBRBMs in modeling the distribution of the spectral envelopes. And both the objective and subjective test results show that the proposed method outperforms the conventional GMM based method.

Index Terms: bandwidth extension, deep neural networks, neural networks, Gaussian-Bernoulli Restricted Boltzmann Machine

1. Introduction

Speech bandwidth expansion (BWE) is a technique that attempts to improve the speech quality by recovering the missing high frequency components using the correlation that exists between the low and high frequency parts of the wide-band speech signal. The BWE techniques have been applied to various tasks, such as speech recognition [1], multicast conference [2], etc. Many approaches have been proposed for speech bandwidth extension during the last decades. Generally, these methods can be classified into two categories: rule-based methods and statistical methods. The rule based methods directly regenerate the high frequency spectral based on the acoustical knowledge of the speech signal, e.g. simply copying a portion of the narrow-band spectrum onto the desired extension frequency components [3]. On the other hand, the statistical methods employ statistical models to estimate the mapping function between the low frequency and high frequency spectral features [4, 5, 6, 7]. By contrast to rule-based methods, statistical methods can construct more precise mapping functions using statistical models. Therefore, statistical methods, especially the GMM-based BWE methods are widely used [5].

Motivated by the success of Deep Neural Networks (DNN) in speech recognition [8], we propose to utilize DNN to estimate a robust mapping function for speech bandwidth extension. Different from the conventional non-linear or linear transformation approaches, the DNN learns both a linear and a non-linear relationship between the low frequency and high frequency spectral envelopes. Thus, DNN can learn a more detailed and precise relationship between the low frequency and high frequency. In our approach, different from the conventional feedforward neural networks for regression tasks, which are usually trained using the back-propagation algorithm under the minimum mean square error criterion, a four-layer DNN is trained layer-by-layer from a cascade of Neural Networks (NNs) and two Gaussian-Bernoulli Restricted Boltzmann Machines (GBRBMs). In the training phase, we first train two exclusive GBRBMs for low frequency and high frequency to obtain the deep networks that capture abstractions for each speech. Then, low frequency feature vectors and high frequency feature vectors are fed into their corresponding GBRBM and high-order features produced by GBRBMs are used to train a concatenating neural network between the two GBRBMs. In the reconstruction phase, the low frequency signal is converted through the trained NNs in the high-order space, and brought back to the cepstrum space using the inverse process of the high frequency GBRBM.

This paper is organized as follows. Section 2 gives an overview of RBM and GBRBM while section 3 explains our speech bandwidth extension method. We show our setup and experimental results in section 4, and section 5 is our conclusion.

2. Preliminaries

Our speech bandwidth extension method uses GBRBM to capture high-order features. We briefly review the GBRBM and its fundamental model, Restricted Boltzmann machine (RBM), in this section.

2.1. RBM

A RBM is a bipartite undirected graphical model. It has a two-layer structure with one visible layer corresponding to a set of visible stochastic variables $\boldsymbol{v} = [v_1, ...v_V]^{\mathrm{T}}$ and one hidden layer corresponding to a set of hidden stochastic variables $\boldsymbol{h} = [h_1, ...h_H]^{\mathrm{T}}$, where V and H denote the number of units in the visible and hidden layers [9]. The joint probability p($\boldsymbol{v}$, $\boldsymbol{h}$) of binary-valued visible units $\boldsymbol{v}$ and binary-valued hidden units $\boldsymbol{h}$ is defined as follows:

$$p(\boldsymbol{v},\ \boldsymbol{h}) = \frac{1}{Z}\exp(-E(\boldsymbol{v},\ \boldsymbol{h})) \tag{1}$$

$$E(\boldsymbol{v},\ \boldsymbol{h}) = -\boldsymbol{a}\boldsymbol{v}^{-T} - \boldsymbol{b}\boldsymbol{h}^{-T} - \boldsymbol{v}^{-T}\boldsymbol{W}\boldsymbol{h} \tag{2}$$

$$Z = \sum_{v,h} \exp(-E(\boldsymbol{v},\ \boldsymbol{h})) \tag{3}$$

where $\boldsymbol{W} = \{w_{ij}\} \epsilon \mathbb{R}^{I \times J}$, $\boldsymbol{a} \epsilon \mathbb{R}^{I \times 1}$ and $\boldsymbol{b} \epsilon \mathbb{R}^{J \times 1}$ are the weight parameter matrix between visible units and hidden units, a bias vector of visible units, and a bias vector of hidden units, respectively.

Because there is no connection between visible units or between hidden units, the conditional probabilities can be written as:

$$p\left(v_j = 1|h\right) = \sigma\left(\boldsymbol{h}^{\mathrm{T}} \boldsymbol{W}_{j:}^{\mathrm{T}} + a_j\right) \tag{4}$$

$$p\left(h_i = 1|v\right) = \sigma\left(\boldsymbol{v}^{\mathrm{T}} \boldsymbol{W}_{:i} + b_i\right) \tag{5}$$

where $\boldsymbol{W_{i:}}$, $\boldsymbol{W_{:j}}$ denote the column vector and the row vector in $\boldsymbol{W}$ respectively, and σ indicates an sigmoid function; i.e. $\sigma\left(x\right) = \frac{1}{1+e^{x}}$.

Conventionally, parameters of a RBM are estimated by maximizing the log-likelihood $\mathcal{L} = \log \prod_n p\left(\boldsymbol{v}_n\right)$. Differentiating partially with respect to each parameter, we obtain

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{ij}} = \left\langle \frac{v_i h_j}{\sigma_i^2} \right\rangle_{data} - \left\langle \frac{v_i h_j}{\sigma_i^2} \right\rangle_{model} \tag{6}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{a}_j} = \left\langle \frac{v_i}{\sigma_i^2} \right\rangle_{data} - \left\langle \frac{v_i}{\sigma_i^2} \right\rangle_{model} \tag{7}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_j} = \left\langle h_j \right\rangle_{data} - \left\langle h_j \right\rangle_{model} \tag{8}$$

where $\left\langle \cdot \right\rangle_{data}$ and $\left\langle \cdot \right\rangle_{model}$ indicate the expectations of the input data and the inner model. Because $\left\langle \cdot \right\rangle_{model}$ is extremely expensive to compute exactly, the contrastive divergence approximation to the gradient is used, where $\left\langle \cdot \right\rangle_{model}$ is replaced by running the Gibbs sampler initialized at the data for one full step [10].

2.2. GBRBM

GBRBM is an extended version of RBM and is suitable for continuous and real-valued data. The units in the visible layer of the GBRBM represent Gaussian stochastic variables, while those in hidden layer represent Bernoulli stochastic variables [11]. The distribution of the stochastic variable described by the GBRBM is defined by an energy function

$$\begin{aligned} E\left(\boldsymbol{v}, \boldsymbol{h}|\Theta\right) = & \sum_{n=1}^{N} \frac{\left(v_n - a_n\right)^2}{2\sigma_n^2} - \sum_{m=1}^{M} b_m h_m \\ & - \sum_{n=1}^{N} \sum_{m=1}^{M} \frac{v_n}{\sigma_n} w_{nm} h_m \end{aligned} \tag{9}$$

where $\Theta = \left(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}\right)$ is the parameter set of an GBRBM, $\boldsymbol{W} \epsilon \mathbb{R}^{N \times M}$ are weights connecting visible and hidden neurons, $\boldsymbol{a} = \left[a_1, ...a_N\right]^{\mathrm{T}}$ and $\boldsymbol{b} = \left[b_1, ...b_M\right]^{\mathrm{T}}$ are the bias terms of visible units and hidden units respectively. σ is the standard deviation associated with a Gaussian visible neuron v_n.

The joint distribution $p\left(\boldsymbol{v}, \boldsymbol{h}\right)$ over $\boldsymbol{v}$ and $\boldsymbol{h}$ is defined by the energy function as

$$p\left(\boldsymbol{v}, \boldsymbol{h}\right) = \frac{1}{Z} \exp\left(-E\left(\boldsymbol{v}, \boldsymbol{h}\right)\right) \tag{10}$$

where

$$Z = \sum_{h} \int \exp(-E(\boldsymbol{v},\ \boldsymbol{h}))d\boldsymbol{v} \tag{11}$$

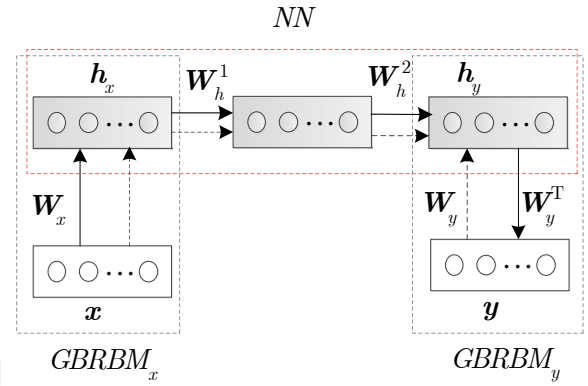

Figure 1: *The structure of the GBRBM BWE system*

The distribution of the visible units is then given as

$$\begin{aligned} p\left(\boldsymbol{v}\right) &= \frac{1}{Z} \sum_{h} \exp\left(-E\left(\boldsymbol{v}, \boldsymbol{h}\right)\right) \\ &= \frac{1}{Z} \exp\left(-\sum_{n=1}^{N} \frac{\left(v_n - a_n\right)^2}{2\sigma_n^2}\right) \\ &\prod_{m=1}^{M} 1 + \exp\left(b_m + \boldsymbol{v}^{\mathrm{T}} \boldsymbol{W}_{:m}\right) \end{aligned} \tag{12}$$

The parameters in GBRBMs can be optimized to maximize the log-likelihood function with a stochastic gradient. Once the parameters are estimated, the conditional probability of $\boldsymbol{h}$ given $\boldsymbol{v}$ and the conditional probability of $\boldsymbol{v}$ given $\boldsymbol{h}$ are respectively written as:

$$p\left(v_i = v|h\right) = \mathcal{N}\left(v; \sigma_i \sum_j h_j w_{ij} + b_i, \sigma_i^2\right) \tag{13}$$

$$p\left(h_i = 1|v\right) = \sigma\left(\sum_i \frac{v_i}{\sigma_i} w_{ij} + b_i\right) \tag{14}$$

where $\mathcal{N}\left(v; \mu, \sigma_i^2\right)$ denotes the probability density function of the Gaussian distribution with mean μ and variance σ_i^2.

3. Speech Bandwidth Extension using DNNs

3.1. Spectral expansion using DNN

Figure 1 shows a flow chart of our method. The proposed model is a four-layer feedforward DNN, including an input layer, two hidden layers and an output layer. In Figure 1, the dashed arrow indicates the training phase while the solid arrow indicates the reconstruction phase. The input and output layers denote the stochastic variables in the spectral vectors of low frequency and high frequency respectively. In the training phase, two GBRBMs are adopted to model the distribution of spectral envelopes for the low frequency and high frequency respectively. Then a NN is employed to model the distribution of the hidden variables extracted from the two GBRBMs. In the reconstruction phase, an input vector of the low frequency is fed to $GBRBM_x$, NN, $GBRBM_y$ in order and then converted to a high frequency vector $\boldsymbol{y}$.

To be more specific, the training process and reconstructing process of the proposed DNN based speech bandwidth extension is conducted as follows:

Step 1 : Train a $GBRBM_x$ using data $\boldsymbol{x}$ of spectral envelopes of the low frequency. Then given the visible samples $\boldsymbol{y}$ and estimated parameters, draw their corresponding hidden samples $\boldsymbol{h_x}$ from the conditional distribution, which is

$$p\left(h_{x,j}=1|x\right)=\sigma\left(\sum_i \frac{x_i}{\sigma_i}w_{ij}+b_{h_x,j}\right) \tag{15}$$

$$h_{x,j}=\sigma\left(\sum_i \frac{x_i}{\sigma_i}w_{ij}+b_{h_x,j}\right) \tag{16}$$

where $\boldsymbol{b_{h_x}}$ are bias vectors of forward inference for low frequency .

Step 2: Train a $GBRBM_y$ using data $\boldsymbol{y}$ of spectral envelopes of the high frequency. Then given the visible samples $\boldsymbol{y}$, draw their corresponding hidden samples $\boldsymbol{h_y}$using mean-field approximation from the conditional distribution, which is

$$p\left(h_{y,i}=1|y\right)=\sigma\left(\sum_j \frac{y_{ji}}{\sigma_j}w_{ij}+b_{h_y,i}\right) \tag{17}$$

$$h_{y,i}=\sigma\left(\sum_j \frac{y_j}{\sigma_j}w_{ij}+b_{h_y,i}\right) \tag{18}$$

where $\boldsymbol{b_{h_y}}$ are bias vectors of forward inference for high frequency.

Step 3: In the last step, a NN is trained, with the projected vectors of the low frequency's acoustic feature $\boldsymbol{h_x}$ being the inputs, and the projected vectors of the corresponding high frequency's feature $\boldsymbol{h_y}$ being outputs. The weight parameters of the NN are estimated to minimize the error between the output $F\left(\boldsymbol{h}_x\right)$ and the target vector $\boldsymbol{h_y}$ as is typical for a NN. Once the weight parameters are estimated, an input vector $\boldsymbol{h_y}$ is converted to

$$\widetilde{\boldsymbol{h}_y}=F\left(\boldsymbol{h}_x\right)=\sigma\left(\boldsymbol{W}_h^2\sigma\left(\boldsymbol{W}_h^1\boldsymbol{h}_x+\boldsymbol{d_1}\right)+\boldsymbol{d_2}\right) \tag{19}$$

where $\boldsymbol{W}_h^1$, $\boldsymbol{W}_h^2$ represents the weight matrices of the first, second layer of the neural network, respectively.

During the reconstruction phase, to map the output $\widetilde{\boldsymbol{h}_y}$ of the NN to the acoustic feature of the high frequency, we just use backward inference of $GBRBM_y$ using Eq. (14), resulting in

$$p\left(\boldsymbol{y}|\widetilde{\boldsymbol{h}_y}\right)=\mathcal{N}\left(\boldsymbol{y};\sigma_y\boldsymbol{W}_y^{\mathrm{T}}\widetilde{\boldsymbol{h}_y}+\boldsymbol{b}_y,\sigma_y^2\right) \tag{20}$$

When minimizing the mean square error (MMSE) estimation rule is adopted for parameter generation, the mapping function takes the form:

$$\begin{aligned}\widetilde{\boldsymbol{y}}_{MMSE}&=E\left\{\boldsymbol{y}|\widetilde{\boldsymbol{h}_y}\right\}\\&=\int_{\Omega_y}\boldsymbol{y}p\left(\boldsymbol{y}|\widetilde{\boldsymbol{h}_y}\right)d\boldsymbol{y}\\&=\int_{\Omega_y}\boldsymbol{y}\mathcal{N}\left(\boldsymbol{y};\sigma_y\boldsymbol{W}_y^{\mathrm{T}}\widetilde{\boldsymbol{h}_y}+\boldsymbol{b}_y\right)d\boldsymbol{y}\\&=\sigma_y\boldsymbol{W}_y^{\mathrm{T}}\widetilde{\boldsymbol{h}_y}+\boldsymbol{b}_y\end{aligned} \tag{21}$$

3.2. Excitation Expansion and power adjustment

Different excitation expansion techniques have been investigated by many researchers and they can be classified into two groups. One is reusing the signal components of the LF excitation signal by spectral folding, spectral translation [12]or nonlinear distortion which includes half-wave rectification [13], full-wave rectification [14], cubic function [15], and the other one is generating new components by noise/sinusoids generator [16] or non-linear processing [17]. Utilizing the LF excitation signal as HF excitation signal results in the best BWE performance in terms of sound quality. Therefore, in this paper, we use this method to predict the HF excitation signal.

It is necessary to adjust the power of the extended excitation signal to the power of the original high frequency excitation signal frame by frame. A codebook mapping method is employed to make the adjustment. To be more specific,

- obtain the energy gain factor $g_1$between the high frequency signal s_h and the low frequency signals_l , which is
$$g_1=\log_{10}\left(\frac{\sum_{n=0}^{N-1}s_h^2\left(n\right)}{\sum_{n=0}^{N-1}s_l^2\left(n\right)}\right) \tag{22}$$
- train and store a codebook C_{g_1} of g_1 using conventional LBG algorithm.
- search an optimal codeword from the codebook C_{g_1}, obtaining the optimal codeword g_1 and the corresponding index i.
- calculate the energy gain factor g_2 between the low frequency signal s_l and s_e obtained from low frequency excitation filtered through high frequency synthesis filter , which is
$$g_2=\log_{10}\left(\frac{\sum_{n=0}^{N-1}s_e^2\left(n\right)}{\sum_{n=0}^{N-1}s_l^2\left(n\right)}\right) \tag{23}$$
- the final gain factor is written as
$$g=\sqrt{\frac{10^{g_i}}{10^{g_2}}} \tag{24}$$

4. Experiments

4.1. Setup

We conducted speech bandwidth extension using one Mandarin Chinese database and an English database. The first Chinese speech database is from the NTT Advanced Technology Corporation (NTT-AT) [18]. The data is sampled at a 16-kHz sampling rate and digitized into 16-bits resolution. The English database is the TIMIT corpus, which also contains 16 kHz speech recordings [19]. A high-pass filtering supplied the high frequency signal. The low frequency signal resulted from a 0.3 to 3.4 kHz band-pass filtering followed by a down-sampling and up-sampling with a factor 2. We use the core training set defined in TIMIT (462 speakers and 4620 utterances) and 64 utterances randomly selected from all speech sound classes in NTT as our training set. The test set consisted of the core test set defined in TIMIT and 32 utterances in NTT.

The baseline system in our experiment was the conversional GMM based BWE. A GMM with 128 components was trained for the baseline system. The 16-order line spectral frequencies (LSFs) [20] were adopted as the spectral feature for the low frequency and high frequency. The frame size and the frame shift

for calculating spectral envelopes was set to 20ms and 10ms respectively. As long as learning of standard deviation is not quite stable, we fixed σ to 1 and normalize the input spectral feature vectors to zero mean and standard deviation 1. The contrastive divergence (CD) learning with 1-step Gibbs sampling was employed to train GBRBMs. The stochastic batch gradient descent algorithm was adopted to update the model parameters. The size of each mini-batch was set to 12 and the learning rate was set to 0.0001. The number of epochs of GBRBMs and NNs were set to 1000 and 300 respectively. The number of hidden units of a GBRBM was fixed to 300.

We investigated on three neural network structures (a 1-layer NN, a 2-layer NN with 1 hidden layer which contains 600 nodes, a 3-layer NN with 2 hidden layers and each hidden layer contains 600 nodes) for the following experiments.

Both objective and subjective measures were used to evaluate the speech bandwidth extension system. The reconstructed speech was measured objectively in terms of distortion between original speech and reconstructed speech. The root mean square log spectral distortion (RMS-LSD) distance in dB and A-B preference tests were used as the objective and subjective measurement, respectively.

4.2. Objective evaluation

We measured the RMS-LSD in the missing high frequency (4-8 kHz). The definition of RMS-LSD [21] is as follows,

$$D\left(A,\widehat{A}\right)=\sqrt{\frac{1}{\omega_2-\omega_1}\int_{-\omega_1}^{\omega_2}\left|20log_{10}\left|\frac{\widehat{A}\left(e^{j\omega}\right)}{A\left(e^{j\omega}\right)}\right|\right|^2 d\omega} \tag{25}$$

where ω_1 and ω_2 are the cut-off frequencies of the missing band; $A\left(e^{j\omega}\right)$and $\widehat{A}\left(e^{j\omega}\right)$ denote the power spectrum of original wideband frame and the power spectrum of corresponding artificially expanded signal respectively. The smaller the value of RMS-LSD is, the closer the reconstructed high frequency to the original high frequency, the better the speech quality is. The RMS-LSD results are shown in Table 1.

Table 1: *RMS-LSD comparison between GMM based BWE and DNNs based BWE.*

Method	RMS-LSD(dB)
GMM based method	8.07
DNNs (1-layer NN)	7.56
DNNs (2-layer NN)	7.29
DNNs (3-layer NN)	7.13

4.3. Subjective evaluation

To evaluate the subjective quality of the proposed DNNs based BWE method, A-B preference tests (A-speech by DNNs based BWE method, B-speech by GMM based BWE method) were carried out and a total of 20 subjects were asked to participate in the preference test. Table 2 shows results of the A-B preference tests.

As shown in Table 2, the proposed method is significantly better than the conventional GMM based BWE method, since DNNs can produce more aurally natural speech than GMM. However, in this paper, only two GBRBMs were studied. In the future, a deeper model which can better describe the nonlinear mapping relationship between low frequency and high frequency will be used by replacing the GBRBMs with deeper stochastic neural networks, such as deep belief networks (DBN).

Table 2: *Subjective preference scores between GMM based BWE and DNNs based BWE.*

Case	Propose method	no preference	GMM-based method
DNNs (1-layer NN)	39	25	36
DNNs (2-layer NN)	42	23	35
DNNs (3-layer NN)	43	23	34

5. Conclusion

In this paper, we proposed a new speech bandwidth extension method using a combination of a low frequency GBRBM, a high frequency GBRBM and concatenating NNs. In our approach, two exclusive GBRBMs for low frequency and high frequency were trained. A NN was then employed to model the joint distribution of the hidden variables extracted from the two GBRBMs. In the reconstruction phase, given a low frequency feature vector, the conditional distribution of the high frequency feature vector can be derived layer-by-layer. Our experimental results showed the efficacy of the proposed method, in comparison to a conventional GMM-based method.

6. References

[1] P. Bauer, J. Abel, V. Fischer, and T. Fingscheidt, "Automatic recognition of wideband telephone speech with limited amount of matched training data," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 22nd European*. IEEE, 2014, pp. 1232–1236.

[2] G. Gandhimathi and S. Jayakumar, "Speech enhancement using an artificial bandwidth extension algorithm in multicast conferencing through cloud services." *Information Technology Journal*, vol. 13, no. 12, 2014.

[3] M. Dietz, L. Liljeryd, K. Kjorling, and O. Kunz, "Spectral band replication, a novel approach in audio coding," in *Audio Engineering Society Convention 112*. Audio Engineering Society, 2002.

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

[5] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using gmm based transformation," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1843–1846.

[6] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using mmse estimation based on a hidden markov model," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–680.

[7] B. Iser and G. Schmidt, "Neural networks versus codebooks in an application for bandwidth extension of speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[8] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[9] 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.

[10] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[11] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14–22, 2012.

[12] N. Enbom and W. B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients," in *Speech Coding Proceedings, 1999 IEEE Workshop on*. IEEE, 1999, pp. 171–173.

[13] J. Epps and W. H. Holmes, "A new technique for wideband enhancement of coded narrowband speech," in *Speech Coding Proceedings, 1999 IEEE Workshop on*. IEEE, 1999, pp. 174–176.

[14] J.-M. Valin and R. Lefebvre, "Bandwidth extension of narrowband speech for low bit-rate wideband coding," in *Speech Coding, 2000. Proceedings. 2000 IEEE Workshop on*. IEEE, 2000, pp. 130–132.

[15] P. J. Patrick and C. Xydeas, "Speech quality enhancement by high frequency band generation," *Digital processing of signals in communications*, pp. 365–373, 1981.

[16] 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, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 3. IEEE, 2006, pp. III–III.

[17] T. Unno and A. McCree, "A robust narrowband to wideband extension system featuring enhanced codebook mapping," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 805–808.

[18] N. A. T. Corporation, "Multi-lingual speech database for telephonometry," http://www.nttat. com/products e/speech, 1994.

[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic phonetic continuous speech corpus cdrom," 1993.

[20] 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, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 665–668.

[21] R. M. Gray, A. Buzo, A. Gray Jr, and Y. Matsuyama, "Distortion measures for speech processing," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 367–376, 1980.