Mapping Neural Networks for Bandwidth Extension of Narrowband Speech

A. Shahina and B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engg.
Indian Institute of Technology Madras,
Chennai - 600 036, India.
{shahina,yegna}@cs.iitm.ernet.in

Abstract

This paper exploits the nonlinear mapping property of feedforward neural networks for estimation of high frequency components (4-8kHz) of the speech signals from the band-limited (0-4kHz) signals. Cepstral coefficients are used to represent the feature vectors of each frame of data. This paper also proposes an approach that uses the autocorrelation method to derive the Linear Prediction (LP) coefficients from the estimated cepstral coefficients that are obtained from the mapping network. This method guarantees the stability of the LP synthesis filter. Informal listenings indicate the effectiveness of the proposed method for estimation of wideband frequency components of speech. The enhanced speech sounds similar to the original wideband speech. Also, it does not contain any distortion that may arise due to spectral discontinuities between adjacent frames.

Index Terms: speech enhancement, bandwidth extension, narrowband speech, mapping neural networks.

1. Introduction

Speech has perceptually significant energy in the 100-8000 Hz range. However when this signal is passed through the analog telephone channel it gets band-limited between 300-3400 Hz range. In the quest for providing high quality speech to the subscribers without modifying the existing telephone network, many researchers have proposed methods to extend the bandwidth of the band-limited telephony speech at the receiver end. The basic idea of these enhancements is to estimate the speech signal components above 3400 Hz and below 300 Hz and to complement the signal in the idle frequency bands with this estimate. Most of the current bandwidth extension schemes use the linearly separable source-filter model of speech production. Here the bandwidth extension is divided into two separate tasks. One is the regeneration of the wideband residual signal, and the other is the recovery of the wideband spectral envelope. The wideband residual is then passed through the wideband Linear Prediction (LP) synthesis filter to produce the wideband speech signal.

In the approaches for the regeneration of wideband spectral envelope, the spectral parameters are trained using a large database of pairs of band-limited and wideband speech sequences. As in speech recognition, the best results are obtained if the training database is as close as possible to the final application [1]. Thus conditions such as recording devices should be chosen carefully. In the linear mapping technique, the band-limited feature vector is mapped to an estimate of the wideband vector by using a linear transformation [2]. The coefficients of the transformation matrix are obtained during training [3]. Improvements in linear mapping using separate matrices to represent voiced sounds, unvoiced sounds, silence, and transition regions in speech was reported in [4]. Codebook mapping technique uses two codebooks, one for narrowband and the other for wideband feature vectors with a one-to-one correspondence between the entries in the two codebooks. For each narrowband test frame, the wideband codevector corresponding to narrowband codevector that is closest to the test vector is chosen as the wideband spectral estimate of that frame. Improvements in codebook mapping was suggested in [5] with the additional use of the phonetic label of each frame during training (supervised clustering). Statistical methods that estimate the wideband parameters based on probabilistic measures, like Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) were reported in [6, 7]. Neural networks were used for bandwidth extension in [8].

This paper exploits the nonlinear mapping property of the Multi-Layered Feedforward Neural Network (MLFFNN) to achieve a mapping of the band-limited spectral features to the wideband spectral features. We propose an approach that uses the autocorrelation method to derive the coefficients of a stable all-pole synthesis filter. The mapping network is shown to capture the relationship between the band-limited and the wideband spectral features. Informal listenings of the enhanced speech show the absence of any distortion due to spectral 'jumps' between adjacent frames of the synthesized speech. The bandwidth extension method discussed in this paper is yet another alternative technique that might help in overcoming a few of the drawbacks of some of the prevailing techniques. This technique is very effective for a speaker-specific case. Future directions would involve ways to extend this technique to the speaker-independent scenario.

This paper is organized as follows: Section 2 describes the proposed approach which is based on the autocorrelation method to derive the spectral features for mapping as well as for constructing the synthesis filter. The nonlinear mapping property of muli-layered feed-forward neural networks is discussed in Section 3. Section 4 explains the experimental details of the proposed approach. Section 5 discusses the performance of the proposed method and suggests future directions.

2. Features for Mapping

Cepstral coefficients are used to represent the feature vector of each frame of data. The cepstral coefficients are derived from the LP coefficients. The cepstral coefficients are obtained from the LP spectrum as follows:

The LP spectrum for a frame of speech is given by

$$|H(k)|^2 = \left| \frac{1}{1+\sum_{n=1}^{p} a_n e^{-j\frac{2\pi}{M}nk}} \right|^2, \quad k = 0, 1, 2 \ldots, M\text{-}1 \quad (1)$$

where a_n are the LP coefficients, and M is the number of spectral values. The inverse Discrete Fourier Transform (DFT) of the log LP spectrum gives the cepstral coefficients c_n. Let

$$S(k) = log|H(k)|^2 \quad (2)$$

Then

$$c_n = \frac{1}{M} \sum_{k=0}^{M-1} S(k) e^{j\frac{2\pi}{M}kn}, \quad n = 0, 1, 2 \ldots, M\text{-}1 \quad (3)$$

Only the first q cepstral coefficients are chosen to represent the LP spectrum. Normally q is chosen much larger than p in order to represent the LP spectrum adequately.

Linearly weighted cepstral coefficients nc_n, n=1, 2, . . . , q, are chosen as a feature vector representing the frame of speech. The linear weighting is done to reduce the sensitivity of the cepstral coefficients to the overall spectral slope [9]. The weighted linear prediction cepstral coefficients (wLPCC) are derived for each frame of band-limited speech and for the corresponding frame of the wideband speech. These pairs of wLPCC vectors are used as input-output pairs to train a neural network model to capture the implicit mapping.

The output of the trained network for each frame of the band-limited data of a test utterance gives an estimate of the wLPCC of the corresponding frame of wideband speech. From these estimated wLPCCs $\hat{c}_n$, $n = 1, 2, \ldots, q$, the estimated log LP spectrum is obtained by using inverse DFT. Let $\hat{S}(k)$, $k = 0, 1, 2, \ldots, M$-1 be the estimated log spectrum. The estimated spectrum $\hat{P}(k)$ is obtained as

$$\hat{P}(k) = e^{\hat{S}(k)}, \quad k = 0, 1, 2 \ldots, M\text{-}1 \quad (4)$$

From the spectrum $\hat{P}(k)$, the autocorrelation function $\hat{R}(n)$ is obtained using inverse DFT of $\hat{P}(k)$.

The first p+1 values of $\hat{R}(n)$ are used in the Levinson-Durbin algorithm to derive the LP coefficients. These LP coefficients for each frame are used to synthesize the enhanced speech by exciting the time varying filter with the regenerated wideband LP residual. The all-pole synthesis filter derived from these LP coefficients is stable because they are derived from the autocorrelation function.

3. Non-linear mapping using MLFFNN

Given a set of input-output pattern pairs, $(\mathbf{a}_l, \mathbf{b}_l)$, l=1, 2, . . . , L the objective of pattern mapping is to capture the implied mapping between the input and output vectors. Once the system behaviour is captured by the network, the network would produce a possible output pattern for a new input pattern not used in the training set. The possible output pattern would be some form of interpolated version of the output patterns corresponding to the input training patterns close to the given test input pattern [10, 11]. The network is said to *generalize* well when the input-output mapping computed by the network is (nearly) correct for the test data that is different from the examples used to train the network [12].

The mapping between the training pattern pairs involves iteratively determining a set of weights $\{w_{ij}\}$ such that the actual output $\mathbf{b}_l^{'}$ is equal (or nearly equal) to the desired output $\mathbf{b}_l$ for all the given L pattern pairs. The weights are determined by using the criterion that the total mean squared error between the desired output and the actual output is to be minimized. The total error $E(W)$ over all the L input-output pattern pairs is given by

$$E(W) = \frac{1}{L} \sum_{l=1}^{L} || \mathbf{b}_l - \mathbf{b}_l^{'} ||^2 \quad (5)$$

$$= \frac{1}{L} \sum_{l=1}^{L} || \mathbf{b}_l - W\mathbf{a}_l ||^2 \quad (6)$$

where W is the weight matrix of the network. To arrive at an optimum set of weights to capture the mapping implicit in the set of input-output pattern pairs, two approaches, namely, gradient descent method and conjugate gradient method are compared. In the gradient descent method the increment in the weight at the $(m+1)^{th}$ iteration is given by

$$\Delta\mathbf{w} = \mathbf{w}(m+1) - \mathbf{w}(m) = -\eta e(m)\mathbf{a}(m) \quad (7)$$

where η is the learning rate parameter, $e(m)$ is the error signal. The weight updation is proportional to the negative gradient of the instantaneous error.

In the conjugate gradient method, the increment in weight at the $(m+1)^{th}$ iteration is given by

$$\Delta\mathbf{w} = \eta(m)\mathbf{d}(m) \quad (8)$$

where the direction of increment $\mathbf{d}(m)$ in the weight is a linear combination of the curent gradient vector and the previous direction of the increment in the weight [10]. The objective is to determine the value of η for which the error $E[\mathbf{w}(m) + \mathbf{d}(m)]$ is minimized for the given values of $\mathbf{w}(m)$ and $\mathbf{d}(m)$. Generally, the conjugate gradient method converges much faster than the gradient descent method.

4. Bandwidth extension of band-limited speech using MLFFNN

4.1. Database

The database for this study comprises of recordings of 6 speakers from the IITM Speech Corpus [13]. A duration of 5 minutes of speech from each speaker is used to obtain the speaker-dependent models. 5 sentences of the speaker are used to test against the model. All the recorded signals are sampled at 16 kHz. The speech signals are downsampled to 8 kHz, and filtered to limit the frequency components upto 3.4 kHz to form the downsampled, band-limited speech signal. This band-limited signal forms the input to the mapping network.

4.2. Features for spectral mapping

LP analysis is performed on both the band-limited and wideband signals. The LP analysis window is 20 msec, with a 10 msec overlap between successive windows. The LP order of the band-limited

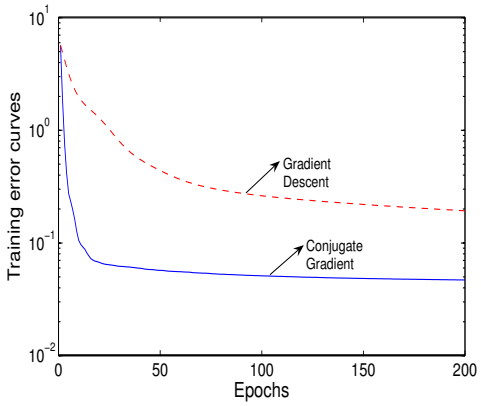

Figure 1: Training error curves for conjugate gradient method and gradient descent method.

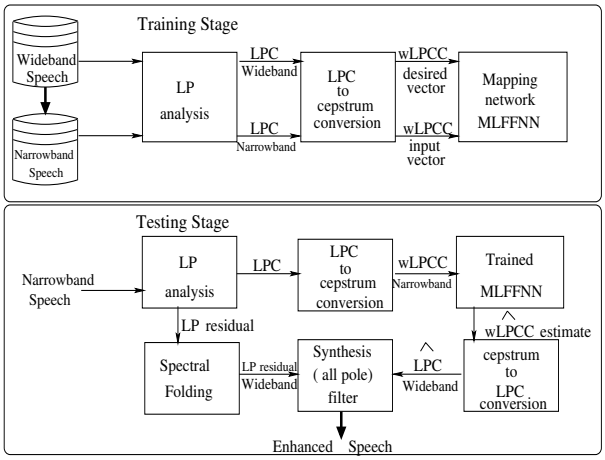

Figure 2: Schematic Diagram of the proposed method for recovery of wideband speech from narrowband speech.

data is 12, and the LP order of the wideband data is 16. The LP coefficients are used to derive 20 dimension wLPCCs as explained in Section 2. Each training pattern is *preprocessed* so that its mean value, averaged over the entire training set, is close to zero. They are then scaled so that they always fall within the range $[-1, 1]$. This accelarates the training process of the network [12]. These *preprocessed* wLPCCs derived from the band-limited signal and the wideband signal form the input-output training pairs, respectively, for the mapping network.

4.3. Training and Testing the mapping network

The structure of the MLFFNN used in this study is $20L30N30N20L$, where L denotes linear units and N denotes nonlinear units. The nonlinear activation function for each unit is given by $\frac{16}{9}tanh(\frac{2x}{3})$, where x is the input activation value. This antisymmetric activation function is suitable for faster learning of the network [12]. The batch mode of training is used here. The weight updation is done using two approaches, namely, the gradient descent method and the conjugate gradient method. In the gradient descent method η is constant, and determines the step size for weight update. In the conjugate gradient method the step size is adjusted at each iteration. A search is made along the conjugate direction to determine the step size that would minimize the performance function along that line. The network is trained for 200 epochs. The conjugate gradient is preferred as it converges much faster than the gradient descent method as can be seen in the training error curves given in Fig. 1.

During testing, the wLPCCs of the band-limited test data is given to the mapping network (trained using the conjugate gradient method). The network produces an output which are the estimated wLPCCs of the corresponding wideband data. The estimated wideband LP coefficients derived from these wLPCCs (as mentioned in Section 2) are used to construct the wideband LP synthesis filter.

4.4. Regeneration of wideband LP residual

The wideband LP residual signal is obtained by spectral folding technique [14]. It is a simple method where the zero-valued samples are added between successive samples of the band-limited (0-3.4 kHz) residual signal. The resulting signal has twice the sampling rate as that of the original. The high frequency spectrum of the modified signal is the mirror image of the band-limited signal. This method is chosen for its simplicity, though better methods could be used. The schematic diagram of the proposed approach is given in Fig. 2.

The regenerated wideband residual signal is passed through the wideband LP synthesis filter to obtain the wideband speech signal. This is passed through a high pass filter and then added to the band-limited signal which is not shown explicitly in the schematic diagram. This is done in order to preserve the low frequency information present in the narrowband input test speech signal.

5. Results and Discussion

Fig. 3 shows the LP spectra of the wideband speech synthesised using the LP coefficients derived from the mapping network, and the spectra of the original wideband speech and the narrowband speech. We see that the LP spectra of the enhanced speech is similar to the LP spectra of the desired wideband speech. This shows that the mapping network has efficiently captured the non-linear mapping in the frequency domain between the band-limited and wideband data.

The spectrograms of the original wideband speech, the band-limited speech and the enhanced speech are shown in Fig. 4. It is seen that the spectral information present in the higher frequency range (above 4000Hz as in fricatives) in the original wideband speech is also present in the enhanced speech. Also the lower frequency information is preserved in the enhanced signal.

Informal listening shows that the enhanced speech is perceptually better than the band-limited speech, and is very similar to the original wideband speech. Also, the enhanced speech has no distortions that may arise due to spectral discontinuities between adjacent frames. This is because the mapping network is able to provide a smooth estimate of the signal. The speech files are available for listening at the website `http://speech.cs.iitm.ernet.in/Main/result/BandwidthExtensionOfSpeechSignals/`. The mapping network has been shown to well capture the mapping for a language-independent, speaker-specific case using downsampled, low pass filtered data. Extending this method for real telephony

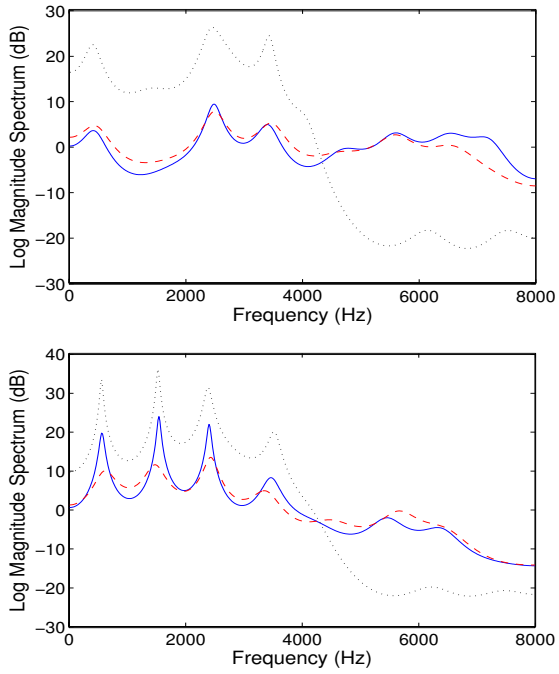

Figure 3: LP spectra of two different segments of the desired wideband speech (solid line), the estimated wideband speech (dashed line), and the narrowband speech (dotted line).

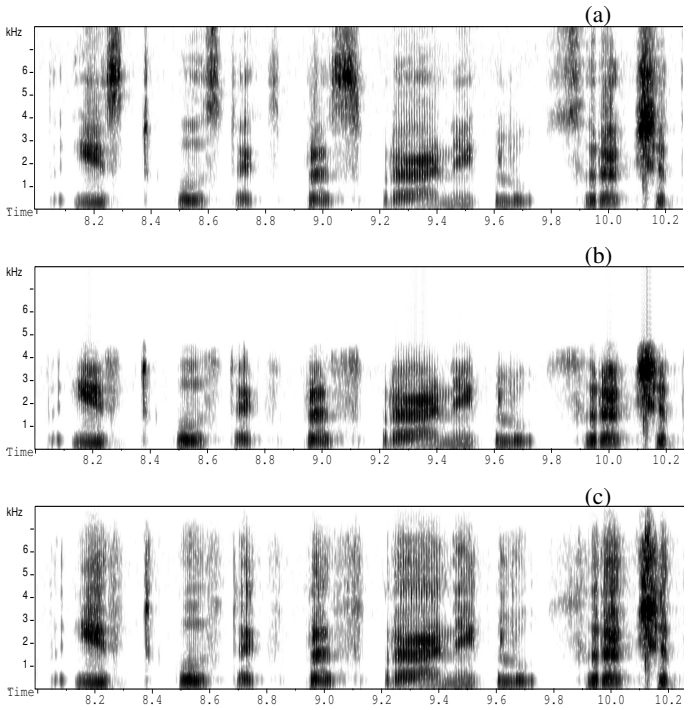

Figure 4: Spectrograms of the (a) wideband speech signal, (b) band-limited speech signal, and (c) the enhanced speech signal for the speech segment *East coast express prayaniglu surakshit* uttered by a male speaker from the Telugu language database of IITM Speech Corpus.

data is currently under progress. This method needs to be extended to a speaker-independent situation, and on real telephone data.

6. References

[1] Bernd Iser, and Gerhard Schmidt, "Bandwidth extension of telephony speech," in *EURASIP Newsletter*, June 2005, vol. 16, pp. 2–24.

[2] Carlos Avendano, Hynek Hermansky and Eric A. Wan, "Beyond Nyquist: Towards the recovery of broad-bandwidth speech from narrow-bandwidth speech," in *Proc. EUROSPEECH* , Madrid, Spain, Sept. 1995, pp. 165–168.

[3] G. Miet, A. Gerrits and J. C. Valire, "Low-band extension of telephone-band speech," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Istanbul, 2000, vol. 3, pp. 1851–1854.

[4] S. Chennoukh, A. Gerrits, G. Miet and R. Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Salt Lake City, Utah, USA, May 2001, vol. 1, pp. 665 – 668.

[5] Rongqiang Hu, Venkatesh Krishnan, and David V. Anderson, "Speech bandwidth extension by improved codebook mapping towards increased phonetic classification," in *Proc. Int. Conf. Spoken Language Processing*, Lisbon, Portugal, Sept. 2005, pp. 1501–1504.

[6] Kim-Youl Park, and Hyung Soon Kim, "Narrowband to wideband conversion of speech using GMM-based transformation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Istanbul, Turkey, June 2000, vol. 3, pp. 1847–1850.

[7] Guo Chen, and Vijay Parsa, "HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Montreal, Quebec, Canada, May 2004, vol. 1, pp. 709–712.

[8] Bernd Iser and Gerhard Schmidt, "Neural networks versus codebooks in an application for bandwidth extension of speech signals," in *Proc. EUROSPEECH* , Geneva, Switzerland, Sept. 2003, pp. 565–568.

[9] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[10] B. Yegnanarayana, *Artificial Neural Networks*, Prentice Hall India, New Delhi, 1999.

[11] Hemant Misra, M. Shajith Ikbal, and B. Yegnanarayana, "Speaker-specific mapping for text-independent speaker recognition," *Speech Comm.*, vol. 39, no. 3-4, pp. 301–310, Feb. 2003.

[12] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice-Hall International, NJ, 1999.

[13] A. Nayeemulla Khan, Suryakanth V. Gangashetty, and S. Rajendran, "Speech database for indian languages - a preliminary study," in *Proc. Int. Conf. Natural Language Processing*, IIT Bombay, India, Dec. 2002, pp. 295–301.

[14] John Makhoul, and Michael Berouti, "High-frequency regeneration in speech coding systems," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Washington DC, Apr. 1979, vol. 4, pp. 428–431.