Bandwidth Expansion of Speech Based on Wavelet Transform Modulus Maxima Vector Mapping

Zhe Chen¹, You-Chi Cheng², Fuiliang Yin¹, Chin-Hui Lee²

¹School of Electronic & Information Engineering, Dalian University of Technology, Dalian, China
²School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, USA
enein@dlut.edu.cn, yccheng@gatech.edu, flyin@dlut.edu.cn, chleece@gatech.edu

Abstract
A novel approach to speech bandwidth expansion based on wavelet transform modulus maxima vector mapping is proposed. By taking advantage of the similarity of the modulus maxima vectors between narrowband and wideband wavelet-analyzed signals, a neural network mapping structure can be established to perform bandwidth expansion given only the narrowband version of speech. Since the proposed algorithm works on the time-domain waveforms it offers a flexibility of variable-length frame selection that facilitates low delay and potentially data-dependent speech segment processing to further improve the speech quality. Evaluations based on both objective and subjective measures show that the proposed bandwidth expansion approach results in high-quality synthesized wideband speech with little perceivable distortion from the original wideband speech signals.

Index Terms: speech bandwidth expansion, wavelet transform modulus maxima, artificial neural network

1. Introduction
It is well-known that wideband speech with 7 KHz bandwidth gives better perceptual quality than narrowband one with 4 KHz bandwidth telephone speech. In recent years a few standardization organizations, such as ITU and 3GPP, present several wideband codec recommendations to promote auditory quality in telecommunication. However these proposals often modify the streaming formats and bit rates in a drastic manner that will result in interoperability and compatibility problems. Bandwidth expansion, also known as bandwidth extension, (BWE) [1] is a way to synthesize wideband speech without altering the streaming formats or bit rates of narrowband speech.

Most of BWE are usually accomplished with recovering spectral envelope which is represented by Fourier transform spectrum, linear prediction coefficients (LPC) or equivalent linear spectrum pair (LSP) and linear spectrum frequency (LSF). The methods of spectral envelope conversion from narrowband to wideband speech are finished by codebook mapping [2]; piecewise linear mapping [3]; Gaussian mixture model mapping [4], hidden Markov model mapping [5], and sparsity-constrained probabilistic mapping [6]. In this paper we propose a new time-domain BWE framework based on wavelet analysis [7]. The wavelet transform modulus maxima (WTMM) [8] obtained in the resulting signals from a multi-resolution representation theory [7], the impulse response of low- and high-pass filter is finished by codebook mapping [2]; piecewise linear mapping [3]; Gaussian mixture model mapping [4], hidden Markov model mapping [5], and sparsity-constrained probabilistic mapping [6]. In this paper we propose a new time-domain BWE framework based on wavelet analysis [7]. The wavelet transform modulus maxima (WTMM) [8] obtained in the resulting signals from a multi-scale wavelet transform, are then used to perform the required waveform mapping to expand input narrowband speech with frequency contents only below 4 KHz to wideband speech with frequency components up to 7 KHz.

2. WTMM-Based Bandwidth Expansion
A system block diagram of WTMM-based bandwidth expansion algorithm is shown in Figure 1. After the input speech s₀(n) is up-sampled to s₀′(n) with 16 KHz sampling rate, wavelet transform, achieved by analysis filter bank, is applied to s₀′(n). Then, the modulus maxima are found in each scale decomposition of s₀′(n). Next, these modulus maxima are converted to its wideband version by WTMM vector mapping. In this study this mapping is accomplished with an artificial neural network. Finally, the output wideband speech signal s₀w(n) with 16 KHz sampling rate can be synthesized from the recovered wideband modulus maxima.

For scale 2^i, the impulse response of low- and high-pass filters, h(i, n) and g(i, n), are defined as follows.

$$h(i, n) = \begin{cases} h(1, n / 2^{i-1}) & n = k \cdot 2^{i-1} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$g(i, n) = \begin{cases} g(1, n / 2^{i-1}) & n = k \cdot 2^{i-1} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$
where \( k = 0, 1, \ldots, 5 \). When the Daubechies wavelet, 'db3' \([9]\), is chosen we have two filters: \( h(1, n) = \{0.23523, 0.57056, 0.32518, -0.09547, -0.06042, 0.02491\} \) and \( g(1, n) = \{0.02491, 0.06042, -0.09547, -0.32513, 0.57056, -0.23523\} \).

The outputs of \( h(i, n) \) and \( g(i, n) \), i.e., \( a(i, n) \) and \( d(i, n) \), known as the approximate and detail signals, can be obtained by convolution operation (denoted as '*') as follows \([9]\):

\[
a(i, n) = h(i, n) * a(i - 1, n) \tag{3}
\]

\[
d(i, n) = g(i, n) * a(i - 1, n) \tag{4}
\]

Similar to locating a maximum in continuous-time signals the modulus maxima of a discrete signal are the points with absolute values larger than its two closest neighbours' and strictly larger than at least one of them \([8]\), i.e., \( x(n_0) \) is a modulus maximum of a discrete signal \( x(n) \) if either condition in Eq. (5) or Eq. (6) is satisfied:

\[
|x(n_0 - 1)| \leq |x(n_0)| \quad \text{and} \quad |x(n_0 + 1)| \leq |x(n_0)| \tag{5}
\]

\[
|x(n_0 - 1)| \leq |x(n_0)| \quad \text{and} \quad |x(n_0 + 1)| \leq |x(n_0)| \tag{6}
\]

where \( | \cdot | \) is the absolute operator.

Since most spectral components of interest in speech are under 4 KHz, in the wavelet analysis of the same scale, the corresponding modulus maxima of detail signals are almost coincident in time between the narrowband and wideband speech. But their amplitudes are different. As an example in Figure 3 the singularity of the original wideband signal at 'O' generates four modulus maxima in the scale 21-24 detail signals \([9]\). Similarly, four modulus maxima of corresponding narrowband signal at 'E-H' in the scale 21-24 detail signals can be identified in Figure 4. Every such set of four modulus maxima forms a 4-dim modulus maximum vector. By transforming the narrowband modulus maxima vector to its corresponding wideband vector with a mapping function the amplitude differences can be compensated.

There are a few algorithms available to synthesize the detail signals from its corresponding modulus maxima \((7-8)\). In this study we simply use a wavelet synthesis filter bank, i.e., the modulus maxima are used directly in synthesis procedure as a detail signal. If a modulus maximum is absent at one time, then the detail signal value is set to 0; Otherwise the detail signal value is set to the corresponding modulus maximum.

The wavelet synthesis filter bank is just a reverse procedure of the wavelet analysis shown in Figure 2. It should be noted that the analysis low- and high-pass filters need to be modified. For example, when the 'db3' wavelet function is used the low- and high-pass filters, in scale 21, should be changed from \( h(1, n) \) and \( g(1, n) \) to \( h(1.5-n) \) and \( g(1.5-n) \), respectively.

The raw modulus maxima values usually have a large variance, and as such not as easy for neural network training. Instead we evaluate the root mean square of the short-time energy over a 65-point rectangle window for smoothing, i.e.,

\[
\text{variance} = \frac{1}{65} \sum_{n=0}^{65} a^2(i, n) \tag{7}
\]

The other key module which needs to be discussed is WTMM vector mapping. Assume that \( v_{wb} = \{v_{wb}(1), v_{wb}(2), v_{wb}(3), v_{wb}(4)\} \) is a 4-dimension modulus maxima vector of narrowband speech, and \( v_{nb} = \{v_{nb}(1), v_{nb}(2), v_{nb}(3), v_{nb}(4)\} \) is its wideband version correspondingly. To extend a narrowband signal, a mapping relationship, \( v_{wb}(i) = f(v_{nb}) \), between modulus maxima vectors must be established. Because the mapping function is often highly nonlinear an artificial neural network (ANN) \([10], [11] \) and \([12]\), is adopted here which is discussed next.

The 3-layer ANN with 1 hidden layer is used for all experiments in this study. We employed 12 units in the hidden layer. The sigmoid nonlinearity is designated as the output function in the hidden layer. The input and output are 4-dim vectors, representing the narrowband and wideband WTMM vectors, respectively, as discussed in the Section 2. When the network is trained, the rule of updating weight and bias values is the Levenberg-Marquardt optimization \([10]\).

3. ANN-Based WTMM Vector Mapping

A 3-layer ANN with 1 hidden layer is used for all experiments in this study. We employed 12 units in the hidden layer. The sigmoid nonlinearity is designated as the output function in the hidden layer. The input and output are 4-dim vectors, representing the narrowband and wideband WTMM vectors, respectively, as discussed in the Section 2. When the network is trained, the rule of updating weight and bias values is the Levenberg-Marquardt optimization \([10]\).

3.1. Training the Networks

Here, we used a series of Matlab functions (Version: R2009a) to train the ANNs. At each epoch a validation set was applied,
and we considered that overtraining was occurring if the validation set error increased over 6 epochs. The mean squared error (MSE) between trained and expected outputs was used to evaluate training performance, as shown in Figure 5.

![Figure 5. MSE of ANN training with back-propagation.](image)

### 3.2. Performance of ANN Mapping

Assume \(v_{ob}\) and \(v_{wb}\) are a pair of 4-dim modulus maximum vectors for narrow- and wide-band speech, and \(v_{wb,\text{map}}\) is an estimate of \(v_{wb}\) obtained by mapping. The second and fourth elements of the modulus maxima vectors, \(v_{ob,\text{map}}(3)\) and \(v_{ob,\text{map}}(4)\), are usually determined by the signals above 4 kHz. They are not our study focus. Thus, we only define \(v_{wb,\text{map}}(1)\) and \(v_{wb,\text{map}}(2)\), corresponding to SNR of the first and second elements of the modulus maxima vectors, for evaluating the mapping performance as follows.

\[
\text{SNR}_{\text{map}}(1) = 10 \log_{10} \left( \frac{\sum v_{ob}(1)^2}{\sum (v_{wb}(1) - v_{wb,\text{map}}(1))^2} \right)
\]

\[
\text{SNR}_{\text{map}}(2) = 10 \log_{10} \left( \frac{\sum v_{ob}(2)^2}{\sum (v_{wb}(2) - v_{wb,\text{map}}(2))^2} \right)
\]

Two sets of vector pairs were used for comparing mapping methods. VP1 is for training, with 8192 vector pairs, and VP2 is for testing, with 2303 vector pairs with different speakers from VP1. We also calculated \(\text{SNR}_{\text{map}}\) of codebook method with a codebook size of 256 [2] to compare with ANN based vector mapping. All results are shown in Table 1. Here it is clear that we can draw a few conclusions: (1) for new speech materials not seen in training codebook mapping generalized not as well as ANN mapping, (2) detailed analysis with the second vector element produces higher SNR values than those done by coarse WTMM analysis as in the first vector element, and ANN was able to achieve an over 30dB SNR for the second modulus maximum, and (3) ANN usually performs better than codebook mapping even with a large codebook size.

### 4. Evaluation of the WBE Algorithm

Three objective measures are adopted to evaluate the quality of the synthesized band-extended speech. They are output signal to noise ratio (\(\text{SNR}_o\)), average log spectrum distance (\(\text{LSD}\)) and visual differences from spectrograms [5].

The output SNR is defined for an \(N\)-sample signal as:

\[
\text{SNR}_o = 10 \log_{10} \left( \frac{\sum_{n=0}^{N-1} [s_{wb}(n)]^2}{\sum_{n=0}^{N-1} [s_{wb}(n) - s_{wb,BWE}(n)]^2} \right)
\]

where \(s_{wb}(n)\) is the original wideband signal, and \(s_{wb,BWE}(n)\) is the recovered wideband signal. As for \(\text{LSD}\) it is defined as

\[
\text{LSD}(k) = \frac{1}{K} \sum_{\omega} \left| \frac{1}{2\pi} \int_{-\pi}^{\pi} |\text{diff}(k, \omega)|^2 d\omega \right|
\]

Here, \(\omega\) is in digital frequency, \(\text{diff}(k, \omega)\) is the log spectrum distance at \(\omega\), \(1/A_s(e^{i\omega})\) and \(1/A_{sb,BWE}(e^{i\omega})\) are the linear predictive all-pole spectra of the original wideband and extended wideband signals, respectively. \(\text{LSD}(k)\) is the \(\text{LSD}\) value of \(k\)-th frame, with \(K\) the total number of frames.

We selected about 10 minutes of speech from the Wall Street Journal corpus for training the mapping ANNs, and we used 6 speakers, three female and three male, from the TIMIT database for testing. Obviously, the training and testing speech are not correlated. The \(\text{SNR}_o\) and \(\text{LSD}\) results are shown in Table 2. We don’t expect the output SNR to be as high as 30dB as in some waveform coders because waveform-based bandwidth expansion will generate quite a bit of noise for high frequency components as computed in Eq. (11). We did observe that the average output SNR value of 15dB in Table 2 is comparable to those obtained in conventional LPC-based speech coding systems. The resulting \(\text{LSD}\) values of 4-5dB are also comparable with results obtained in other studies when speech of different talkers from the training speakers are tested [6]. As for subjective listening test the proposed algorithm yields very good results in most situations.

### Table 2. Evaluation of WBE synthesized speech.

<table>
<thead>
<tr>
<th>TIMIT Talker</th>
<th>LSD</th>
<th>(\text{SNR}_o)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female 1</td>
<td>5.11 dB</td>
<td>14.92 dB</td>
</tr>
<tr>
<td>Female 2</td>
<td>4.14 dB</td>
<td>14.62 dB</td>
</tr>
<tr>
<td>Female 3</td>
<td>4.84 dB</td>
<td>15.19 dB</td>
</tr>
<tr>
<td>Male 1</td>
<td>5.36 dB</td>
<td>14.90 dB</td>
</tr>
<tr>
<td>Male 2</td>
<td>5.19 dB</td>
<td>14.43 dB</td>
</tr>
<tr>
<td>Male 3</td>
<td>4.80 dB</td>
<td>16.03 dB</td>
</tr>
</tbody>
</table>

The spectrograms of an analyzed segment from Female 2 for the original wideband, the synthesized wideband, and synthesized noisy speech signals are shown in Figures 6-8. To generate Figure 8 the narrowband speech segment is corrupted with white noise artificially at \(\text{SNR}=10\)dB.

Although there are some visible discrepancies in the high frequency section of the spectrograms between Figures 6 and
subjective listening tests gave little perceptual difference in this case. As for the effect of bandwidth expansion on noisy speech in Figure 8 we could not observe an over-emphasis of noise above 4 KHz. Such a mismatch can be perceptually annoying as in some BWE systems with codebook mapping.

Finally the proposed algorithm can be performed on a sample-by-sample basis, so its frame length can be variable to match any coder performance requirement. Algorithm delay comes mainly from the wavelet filter operations while ANN mapping consumes most of the computation cycles.

5. Summary

We have presented a waveform-based speech bandwidth expansion framework by exploring the similarity between the wavelet transform modulus maxima of the corresponding narrowband and wideband speech signals. An ANN can then be established to predict the missing modulus maxima of corresponding wideband speech given the modulus maxima of narrowband speech. One key advantage is that the proposed algorithm works for any variable-length speech segments. This nice property facilitates low-delay bandwidth expansion, and results in a good quality for synthesized wideband speech.

6. Acknowledgment

The authors thank Chengyuan Ma of Georgia Tech for sharing his speech tools. The first author also thanks China Scholarship Council for supporting his visiting stay at Georgia Institute of Technology from August 2010 to August 2011. This work was supported in part by the Dalian Municipal Science and Technology Fund Scheme No.2008J23JH025, National Natural Science Foundation of China No.60772161, No.60372082 and Specialized Research Fund for the Doctoral Program of Higher Education No. 200801410015.

7. References