16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

ARTIFICIAL BANDWIDTH EXTENSION OF NARROW-BAND SPEECH SIGNALS VIA HIGH-BAND ENERGY ESTIMATION

Tenkasi Ramabadran and Mark Jasiuk

Motorola Labs, Motorola Inc.,
1301 East Algonquin Road, Schaumburg, IL 60196, USA
phone: + (1) 847-576-3723, fax: + (1) 847-576-6030, email: Tenkasi.Ramabadran@.motorola.com
Mark.Jasiuk@motorola.com

ABSTRACT

In this paper, we describe a novel method of tackling the problem of artificially extending the bandwidth of a narrow-band speech signal. For a given narrow-band signal, we first estimate the energy in the high-band. The high-band energy is then used to select a suitable high-band spectral envelope shape that is consistent with the estimated high-band energy while simultaneously ensuring that the resulting wide-band spectral envelope is continuous at the boundary between narrow-band and high-band. The scalar high-band energy parameter thus effectively controls the artificial information added to the high-band of the bandwidth extended output speech signal. Artifacts in the output speech are minimized by adapting the high-band energy parameter appropriately. Formal subjective listening tests show that the bandwidth extended speech output generated by the described method outscores the input narrow-band speech by 0.25 MOS.

1. INTRODUCTION

The acoustic bandwidth of speech signals in most of today's telephone communication systems is limited to around 300 – 3400 Hz, the so-called "narrow-band". This limitation in frequency range, originating from former analogue transmission techniques, is the main reason for the muffled quality and reduced intelligibility of telephone speech as compared to natural speech. On the other hand, wideband speech, typically defined by the frequency range of 50 – 7000 Hz, sounds more natural and has higher intelligibility than narrow-band speech. Telephone communication systems capable of transmitting wideband speech signals are expected to be deployed in the future as evidenced by the fact that speech coding schemes have been developed and standardized for the wider bandwidth [1] [2]. However, such deployment will likely be gradual because of economic reasons. In the meantime, artificial bandwidth extension (BWE) techniques that seek to extend the perceived acoustic bandwidth of an input narrow-band speech signal by adding synthesized signals to the high-band (e.g., 3400 – 7000 Hz) and occasionally low-band (e.g., 100 – 300 Hz) provide an attractive alternative. The bandwidth extended speech can potentially provide better quality and higher intelligibility than the narrow-band speech. The added signals are synthesized based only on the available narrow-band information, and so no increase in transmission bit rate is necessary. Furthermore, bandwidth extension is implemented at the receiver, which is hence the only part of the communication system that needs to be modified.

A number of techniques [3] – [7] have been proposed over the years for bandwidth extension of narrow-band (NB) speech. Most of these techniques are based on a parametric (viz., *source-filter*) model of speech production whereby the speech signal is regarded as an excitation *source* signal that has been acoustically *filtered* by the vocal tract. In a typical parametric BWE technique, the input NB speech is first analyzed to extract the spectral envelope information and the residual excitation information via linear predictive (LP) analysis. From the narrow-band excitation signal, the wideband excitation signal is estimated. Similarly, from the narrow-band envelope, the wideband envelope is estimated. The estimated wideband excitation and envelope are combined in an LP synthesis filter to generate estimated wideband speech. The high-band portion of the estimated wideband speech is extracted using a high-pass filter (HPF), adjusted for gain, and combined with the input NB speech to generate the bandwidth extended speech. The various parametric techniques reported in the literature differ mostly in the way the wideband envelope is estimated and sometimes in the way the wideband excitation is estimated.

While BWE speech that sounds like wideband speech can be generated using any of the reported techniques, the main obstacle to the commercialization and widespread use of BWE technology is the presence of objectionable artifacts in the output speech that degrades its quality. It is known that overestimation of high-band energy is a source of artifacts [6]. In the technique described in this paper, therefore, the estimation of high-band energy and its adaptation play a critical role in minimizing artifacts and generating high-quality BWE speech. From the input narrow-band signal, the high-band energy is first estimated. The estimated high-band energy is then used to select a high-band spectral envelope that is consistent with the estimated energy while simultaneously ensuring that the resulting wideband spectral envelope is continuous at the boundary between narrow-band and high-band. The scalar high-band energy parameter

thus effectively controls the information added to the high-band of the BWE speech. Artifacts are minimized by adapting this parameter appropriately depending on estimation accuracy and/or narrow-band signal characteristics thereby enhancing BWE speech quality. The paper is structured as follows. The overall BWE system block diagram is discussed in Section 2. In Section 3, some of the design details are described. Experimental results are presented in Section 4. Finally, in Section 5, our conclusions are provided.

2. SYSTEM BLOCK DIAGRAM

Figure 1 shows the system block diagram. The input narrow-band speech sampled at 8 kHz is fed into the system at top left. Processing of the input NB speech is performed on a frame-by-frame basis, where a frame is defined as a sequence of N consecutive samples over a duration of T sec. Frame durations typically range from 10 to 30 ms. Consecutive frames may overlap each other, e.g., by 50%. The input NB speech is first up-sampled by a factor of 2, i.e., to 16 kHz, to generate up-sampled narrow-band speech. Linear predictive (LP) analysis is performed on the input NB speech to extract LP coefficients $\{1, a_1, a_2, \ldots, a_P\}$ modelling the NB spectral envelope, where the model order P is typically 10. These coefficients are interpolated by a factor of 2 (by inserting a zero between every pair of coefficients) and then used to analyze (i.e., inverse filter) the up-sampled NB speech to generate the NB residual excitation at 16 kHz. The NB residual excitation is full-wave rectified (FWR) to extend its bandwidth to the entire band (0 – 8 kHz) through the *non-linear* rectification operation and high-pass filtered (HPF) to obtain the high-band (HB) residual excitation. The bandwidth of the HB residual excitation, e.g., is 3400 – 8000 Hz. High-band noise excitation is separately generated by high-pass filtering a pseudo-random noise sequence.

The HB residual excitation and the HB noise excitation are combined in a mixer according to a voicing level v provided by the Estimation and Control Module (ECM) shown at the right. Inputs to and outputs from the ECM are shown by dashed lines. Inputs to the ECM are the input NB speech, the up-sampled NB speech, and the LP coefficients modelling the NB spectral envelope. Outputs from the ECM are the voicing level v, the high-band energy E_{hb}, and the wide-band spectral envelope SE_{wb}. The voicing level v ranges from 0 for unvoiced speech to 1 for fully voiced speech. When the voicing level is 0, the mixer outputs only HB noise excitation; when the voicing level is 1, the mixer outputs only HB residual excitation; and when the voicing level is somewhere in between the two bounds corresponding to mixed-voiced speech, the mixer outputs a suitable combination of HB noise excitation and HB residual excitation. The mixer output is henceforth referred to as the high-band (HB) excitation. The HB excitation is scaled to the energy level E_{hb} and combined with the up-sampled NB speech to form a zeroth approximation of the wideband speech. This signal is then filtered by the equalizer filter, which imposes the wide-band spectral envelope SE_{wb} provided by ECM onto the

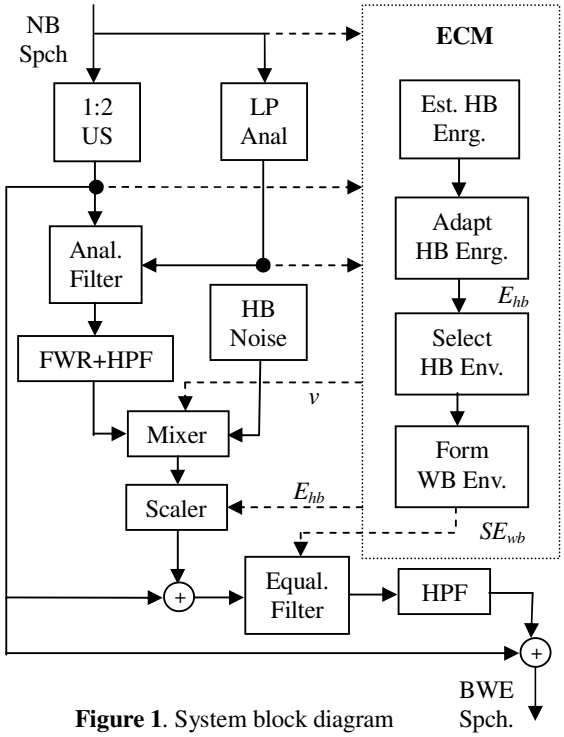

Figure 1. System block diagram

input signal to estimate a better approximation of the wide-band speech. The estimated wideband speech is high-pass filtered to extract the high-band (HB) speech. The HB speech and the up-sampled NB speech are combined together to form the output BWE speech shown at bottom right. Optionally, a bass-boost filter can be used to recover some of the missing low frequency (e.g., 100 – 300 Hz) information in the up-sampled NB speech before it is added to HB speech to form the output.

Within the ECM, the high-band energy is first estimated from the available narrow-band information. The estimated high-band energy is then adapted to minimize the artifacts in the BWE speech. Using the adapted high-band energy E_{hb}, an appropriate high-band envelope is selected. The high-band envelope is combined with the narrow-band envelope to form the estimated wideband envelope SE_{wb}. The blocks within the ECM will be described in greater detail in Section 3.

The HB excitation is obtained by mixing the HB residual excitation and the HB noise excitation as described earlier. For a voiced speech frame, the NB residual excitation is voiced, and when it is processed by the "FWR+HPF" block and the "mixer" block, the harmonic structure is still retained in the HB excitation spectrum. For an unvoiced speech frame, the HB noise excitation provides a noise-like spectrum for the HB excitation. For a mixed-voiced frame, the spectrum of the HB excitation has both harmonic and noise-like structures. This approach to generating the HB excitation results in a natural-sounding BWE speech. In generating the estimated wideband speech, an equalizer filter is used instead

of the traditional LP synthesis filter. The equalizer filter uses an overlap-add (OLA) analysis and synthesis approach [8] for its operation. Raised cosine windows with perfect reconstruction property and 50% overlap are used for this purpose. For a given (windowed) input frame, the equalizer filter determines its spectral envelope SE_{in}, e.g., using LP analysis. The target envelope SE_{wb} is provided by the ECM. The equalizer filter magnitude response is then computed as $SE_{wb}(\omega)/SE_{in}(\omega)$, where ω is the normalized frequency in radians/sample, and its phase response is set to zero. The equalizer filter thus attempts to impose the desired spectral envelope shape onto the input signal. The equalizer filter offers several advantages: (a) since the phase response of the equalizer filter is zero, the different frequency components of its output are time-aligned with corresponding frequency components of its input; this can be useful, e.g., for voiced speech, because high energy segments (e.g., glottal pulses) of the HB excitation will be time-aligned with and hence masked by the corresponding high energy pitch pulses of the up-sampled NB speech, (b) the equalizer filter response is specified in the frequency domain, so a better and finer control over different parts of the spectrum is possible, (c) the input to the equalizer filter does not need to have a flat spectrum, and (d) iterations are possible to improve the effectiveness of the filter at the cost of additional delay and complexity; that is, the filter output can be fed back into the input to be equalized again and thereby improve filter performance. The equalizer filter described here is similar in principle to the filter bank equalizer used in the G.729.1 standard [9].

3. DESIGN DETAILS

The design details of the different blocks within the ECM are described below.

3.1 Estimation of High-Band Energy

In previous approaches, the high-band energy is usually estimated in terms of the narrow-band energy, typically as a ratio. Here, we estimate the high-band energy in terms of a transition-band energy, where the transition-band is defined as a frequency band contained within the narrow-band and close to the high-band, i.e., it serves as a transition to the high-band, e.g., 2500 – 3400 Hz. Intuitively, one would expect the transition-band to be better correlated with the high-band than the entire narrow-band, which is borne out in experiments. Denoting the transition-band energy as E_{tb} (in dB), the high-band energy E_{hb0} (in dB) is estimated as

$$E_{hb0} = \alpha E_{tb} + \beta$$

where the coefficients α and β are chosen to minimize the mean squared error between the true and estimated high-band energy values over a large number of frames from a training database. Estimation accuracy is further improved by using contextual information provided by additional parameters derived from available narrow-band information. These parameters are: (a) normalized zero-crossing parameter *zc* (range: 0 – 1) computed from the input NB speech, (b) spectral flatness measure parameter *sfm* (range: 0 – 1) computed from the spectral envelope of the up-sampled NB speech within the 300 – 3400 Hz band as the ratio of the geometric mean to the arithmetic mean, and (c) transition-band spectral envelope shape parameter *tbs* computed from the spectral envelope shape of the up-sampled NB speech using a Vector Quantizer (VQ) codebook of 64 shapes designed using the training database. The three dimensional *zc-sfm-tbs* parameter space is partitioned as follows. The *zc-sfm* plane is partitioned into 12 regions thereby giving rise to possibly 12 × 64 = 768 regions in the three dimensional space. Out of these, only about 500 regions have sufficient data points from the training database, and so for each of these about 500 regions, separate sets of α and β coefficients are selected. Even further improvement in estimation accuracy is achieved by increasing the order of the estimator, e.g., as

$$E_{hb0} = \alpha_3 E_{tb}^3 + \alpha_2 E_{tb}^2 + \alpha_1 E_{tb} + \beta .$$

In this case, different sets of α_3, α_2, α_1, and β coefficients are selected for each of the about 500 regions.

3.2 Adaptation of High-Band Energy

The estimated high-band energy is adapted as described below to minimize artifacts and thereby enhance the quality of the output BWE speech. Estimation of high-band energy is prone to errors. Since over-estimation leads to artifacts, the estimated high-band energy is biased to be lower by an amount proportional to the standard deviation of the estimation error as

$$E_{hb1} = E_{hb0} - \lambda \cdot \sigma$$

where E_{hb1} is the adapted high-band energy in dB, $\lambda \geq 0$ is a proportionality factor, and σ is the standard deviation of the estimation error in dB. By "biasing down" the estimated high-band energy as above, the probability (or number of occurrences) of energy over-estimation is reduced, thereby reducing the number of artifacts. Also, the amount by which the estimated energy is reduced is proportional to how good the estimate is – a more reliable (i.e., low σ value) estimate is reduced by a smaller amount than a less reliable estimate. While designing the high-band energy estimator, the σ value corresponding to each partition of the *zc-sfm-tbs* parameter space is computed from the training speech database and stored for later use. This "bias down" of estimated energy has an added benefit for voiced frames – that of masking any "noisy" artifacts arising from errors in high-band spectral envelope shape estimation. However, for unvoiced frames, if the reduction in the estimated high-band energy is too high, the output BWE speech no longer sounds like wideband speech. To counter this, the estimated high-band energy is further adapted depending on the voicing level *v* as

$$E_{hb2} = E_{hb1} + (1\text{-}v) \cdot \delta_1 + v \cdot \delta_2$$

where E_{hb2} is the voicing-level adapted high-band energy in dB and δ_1 & δ_2 ($\delta_1 > \delta_2$) are constants in dB. The choice of δ_1 and δ_2 depends on the value of λ used for the "bias down" and are determined empirically to yield the best-sounding output speech. The voicing level *v* itself is estimated from the normalized zero-crossing parameter *zc* and

two thresholds ZC_{low} and ZC_{high}. If zc is below ZC_{low}, v is 1; if zc is above ZC_{high}, v is 0; otherwise, the range between ZC_{low} and ZC_{high} is linearly mapped onto the range 0 to 1 for v. Occasionally, there are frames for which the high-band energy is grossly under- or over-estimated, the so called outliers. Such errors are reduced by smoothing the estimate using, e.g., a three-point averaging filter as

$$E_{hb3} = [E_{hb2}(k\text{-}1) + E_{hb2}(k) + E_{hb2}(k\text{+}1)] / 3$$

where E_{hb3} is the smoothed estimate and k is the frame index. The smoothed energy estimate E_{hb3} is further adapted depending on whether the frame is *steady-state* or *transient*. A frame is considered steady-state if it is close to both of its neighboring frames in a spectral sense (using the Itakura distance measure, for example) as well as in terms of energy; otherwise, it is transient. A steady state frame is able to mask errors in high-band energy estimation much better than transient frames. Accordingly, the smoothed energy estimate is further adapted as

$$\begin{aligned} E_{hb4} &= E_{hb3} + \mu_1 && \text{for steady-state frames} \\ E_{hb4} &= \min(E_{hb3} - \mu_2, E_{hb2}) && \text{for transition frames} \end{aligned}$$

where $\mu_2 > \mu_1 \geq 0$, are empirically chosen constants in dB to achieve good output speech quality. Finally, the estimated high-band energy is adapted depending on the occurrence of an onset/plosive. An onset/plosive presents a special problem because of the following reasons: (a) estimation of high-band energy near an onset/plosive is difficult, (b) pre-echo type artifacts may occur in the output speech because of the typical block processing employed, and (c) plosive sounds (e.g., [p], [t], and [k]), after their initial energy burst, have characteristics similar to certain sibilants (e.g., [s], [ʃ], and [ʒ]) in the narrow-band but quite different in the high-band leading to energy over-estimation and consequent artifacts. An onset/plosive is detected at the current frame if the input NB speech energy of the preceding frame is below a certain threshold and the energy difference between the current and preceding frames exceeds another threshold. High-band energy adaptation upon detection of an onset/plosive is done as follows:

$$\begin{aligned} E_{hb}(k) &= E_{min} && \text{for } k = 1, \ldots, K_{min} \\ E_{hb}(k) &= E_{hb4}(k) - \Delta && \text{for } k = K_{min}\text{+}1, \ldots, K_T \\ E_{hb}(\mathrm{k}) &= E_{hb4}(\mathrm{k}) - \Delta + \Delta_T(k\text{-}K_T) && \text{for } k = K_T\text{+}1, \ldots, K_{\max} \end{aligned}$$

For the first K_{min} frames starting with the frame (k = 1) at which the onset/plosive is detected, the high-band energy is set to the lowest possible value E_{min}. For the subsequent frames (i.e., for $k = K_{min}$+1 to K_{max}), energy adaptation is done only as long as the voicing level $v(k)$ of the frame exceeds a threshold V_1. Whenever the voicing level of a frame within this range becomes less than or equal to V_1, the onset/plosive energy adaptation is immediately stopped. This feature enforces a shorter duration of energy adaptation for certain sounds, e.g., voiced onsets. If the voicing level $v(k)$ is greater than V_1, then for $k = K_{min}$ + 1 to $k = K_T$, the high-band energy is decreased by a fixed amount Δ. For $k = K_T$ + 1 to $k = K_{max}$, the high-band energy is gradually increased from $E_{hb4}(k) - \Delta$ towards $E_{hb4}(k)$ by means of the pre-specified sequence $\Delta_T(k\text{-}K_T)$ and at $k = K_{max}$ + 1, $E_{hb}(k)$ is set equal to $E_{hb4}(k)$. If no onset/plosive is detected, the final adapted high-band energy estimate E_{hb} is set equal to E_{hb4}.

3.3 Selection of High-Band Spectral Envelope Shape

To select a high-band spectral envelope shape corresponding to a given high-band energy, we proceed as follows. Starting with a large training database of wide-band speech sampled at 16 kHz, the wide-band spectral magnitude envelope is computed for each speech frame using standard LP analysis or other techniques. From the wide-band spectral envelope of each frame, the high-band portion corresponding to 3400 – 8000 Hz is extracted and normalized by dividing through by the spectral magnitude at 3400 Hz. The resulting high-band spectral envelopes have thus a magnitude of 0 dB at 3400 Hz. The high-band energy corresponding to each normalized high-band envelope is computed next. The collection of high-band spectral envelopes is then partitioned based on the high-band energy, e.g., a sequence of nominal energy values differing by 1 dB is selected to cover the entire range, and all envelopes with energy within 0.5 dB of a nominal value are grouped together. For each group thus formed, the average high-band spectral envelope shape is computed and subsequently the corresponding high-band energy. In Figure 2, a set of 60 high-band spectral envelope shapes at different energy levels is shown. Counting from the bottom, the 1st, 10th, 20th, 30th, 40th, 50th, and 60th shapes (referred to henceforth as pre-computed shapes) were obtained using a technique similar to the one described above. The remaining 53 shapes were obtained by simple linear interpolation (in the dB domain) between the nearest pre-computed shapes. The energies of these shapes range from about 4.5 dB for the 1st shape to about 43.5 dB for the 60th shape with an average energy resolution of about 0.65 dB. Given the high-band energy for a frame, it is then a simple matter to select the closest matching high-band spectral envelope shape. It is seen from Figure 2 that small changes in high-band energy correspond to small changes in high-band spectral envelope shapes. This permits the explicit control of the time evolution of the high-band spectral envelope shape by controlling the time evolution of the high-band energy. Smooth evolution of the high-band spectrum, at least within distinct speech segments, can be important for ensuring natural-sounding, high-quality output BWE speech.

3.4 Formation of Wideband Spectral Envelope

Using the technique described above for the selection of the high-band spectral envelope shape, the wideband spectral envelope SE_{wb} is formed as follows. From the up-sampled NB speech frame, the narrow-band magnitude spectral envelope SE_{nb} is computed and its value at 3400 Hz is determined. Let this value in dB be denoted as M_{3400}. Given the adapted high-band energy E_{hb} in dB, we select the high-band spectral envelope shape that is closest in energy to E_{hb}-M_{3400}. Let this shape be denoted as $SE_{closest}$. The high-band spectral envelope SE_{hb} is then given by M_{3400} + $SE_{closest}$. The envelopes SE_{nb} and SE_{hb} are then spliced to form SE_{wb}. It is clear that the wideband spectral envelope SE_{wb} formed using the

above procedure is continuous at the junction between the narrow-band and high-band. It also has the correct high-band energy, viz., E_{hb}.

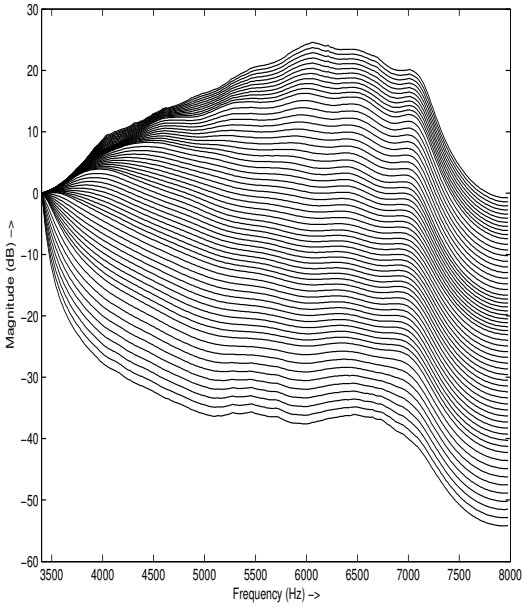

Figure 2. High-band spectral envelope shapes at different high-band energy levels

4. EXPERIMENTAL RESULTS

A formal subjective listening test was conducted to evaluate the quality of the BWE speech generated by the system described in this paper. A bass-boost filter was used to recover some of the low-frequency information in the BWE speech. The speech material used in the test consisted of 32 Harvard sentence pairs spoken by 4 males and 4 females with 4 sentence pairs each. Besides the original WB speech (0 – 8000 Hz), NB speech (300 – 3400 Hz), and BWE speech, several other processed speech conditions were included in the test. For example, filtered speech data with different bandwidths bounded by the bandwidths of NB speech and WB speech were included. MNRU (Modulated Noise Reference Unit) conditions ranging from 6 dB to 42 dB were included. The speech material was presented to a group of 32 listeners monaurally using Sennheiser HD 25-1 headphones at a sound level of 79 dB-SPL. The listeners were asked to grade each sentence pair on a scale of 1 to 5 (1 – bad, 2 – poor, 3 – average, 4 – good, and 5 – excellent). A total of 256 votes was collected for each tested condition and the mean opinion score (MOS) was calculated by averaging these votes. Some of the MOS results are presented in Table 1. It is seen that the BWE speech outscores the input NB speech by 0.25 MOS. The 95% confidence interval for the results is approximately ± 0.1 MOS.

Table 1. Subjective listening test results

Test Condition	MOS
WB speech (0 – 8000 Hz)	4.33
Filtered speech (0 – 6000 Hz)	4.27
Filtered speech (300 – 6000 Hz)	4.00
Filtered speech (0 – 4000 Hz)	4.04
Filtered speech (300 – 4000 Hz)	3.82
Filtered speech (0 – 3400 Hz)	3.68
NB speech (300 – 3400 Hz)	3.64
BWE speech (150 – 8000 Hz)	3.89

5. CONCLUSIONS

A bandwidth extension system with several novel features was described. The main feature of the system is to estimate the high-band energy accurately and select the high-band envelope shape based on this energy. A single parameter thus controls the high-band information added and this parameter is adapted to minimize artifacts in the output BWE speech. The BWE speech is clearly preferred by the listeners over the input NB speech. Future research will explore methods to enhance high-band spectral envelope shape estimation. Besides high-band energy, other parameters derived from input NB speech can perhaps be used to achieve a better selection of the high-band spectral envelope shape. Reduction of delay and complexity of the method and improved energy estimation are also subjects of future research.

REFERENCES

[1] B. Bessette, et al., "The Adaptive Multirate Wideband Speech Codec," *IEEE Transaction on Speech and Audio Processing*, Vol. 10, No. 8, pp. 620-636, November 2002.

[2] V. Krishnan, et al., "EVRC-Wideband: The New 3GPP2 Wideband Vocoder Standard," in *Proc. ICASSP 2007*, Honolulu, Hawai'i, USA, April 15-20, 2007, pp. II-333–II-336.

[3] Y. M. Cheng, et al., "Statistical Recovery of Wideband Speech from Narrowband Speech," *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 4, pp. 544-548, October 1994.

[4] H. Carl and U. Heute, "Bandwidth Enhancement of Narrow-Band Speech Signals," in *SIGNAL PROCESSING VII*: *Theories and Applications, EUSIPCO 1994*, pp. 1178-1181.

[5] J. Epps, "Wideband Extension of Narrowband Speech for Enhancement and Coding," *Ph.D. Thesis*, School of Electrical Engineering and Telecommunications, The University of New South Wales, September 2000.

[6] M. Nilsson and W.B. Kleijn, "Avoiding Over-Estimation in Bandwidth Extension of Telephony Speech," in *Proc. ICASSP 2001*, Salt Lake City, Utah, USA, May 7-11, 2001, pp. 869-872.

[7] J. Kontio, L. Laaksonen, and P. Alku, "Neural Network-Based Artificial Bandwidth Expansion of Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 3, pp. 873-881, March 2007.

[8] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[9] B.Geiser, et.al., "Bandwidth Extension for Hierarchical Speech and Audio Coding in ITU-T Rec. G.729.1," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2496-2509, November 2007.