BANDWIDTH EXTENSION OF TELEPHONE SPEECH USING A FILTER BANK IMPLEMENTATION FOR HIGHBAND MEL SPECTRUM

Hannu Pulakka[1], *Ville Myllylä*[2], *Laura Laaksonen*[2], *Paavo Alku*[1]

[1]Aalto University School of Science and Technology, Department of Signal Processing and Acoustics, Finland
[2]Nokia Devices, Audio Technology R&D, Finland
email: hannu.pulakka@tkk.fi

ABSTRACT

The limited audio bandwidth used in telephone systems degrades both the quality and the intelligibility of speech. This paper presents a new method for the bandwidth extension of telephone speech. Frequency components are added to the frequency band 4–8 kHz using only the information in the narrowband speech. First, a wideband excitation is generated by spectral folding from the narrowband linear prediction residual. The highband of this signal is divided into four subbands with a filter bank, and a neural network is used to weight the subbands based on features calculated from the narrowband speech. Bandwidth-extended speech is obtained by summing the weighted subbands and the original narrowband signal. Listening tests show that this new method improves speech quality compared with a previously published bandwidth extension method.

1. INTRODUCTION

Most telephone systems in use today transmit narrowband speech using the traditional telephone band of 300–3400 Hz or only a slightly wider audio bandwidth. For example, the Adaptive Multi-Rate (AMR) codec, which is widely used in the GSM system, is a narrowband speech codec. The narrow bandwidth degrades speech quality, naturalness, and intelligibility. Significant improvement can be achieved by wideband speech coding systems. For example, the AMR-WB speech codec transmits the audio bandwidth of 50–7000 Hz. AMR-WB has been selected as the wideband codec to be used in GSM and in the third-generation (3G) mobile communication system, but the transition period from narrowband to wideband telephony is expected to take a long time.

The quality difference between narrowband and wideband speech can be reduced by artificial bandwidth extension (ABE) of narrowband speech. This refers to techniques that artificially generate frequency components in spectral regions that are not available in the narrowband signal. Bandwidth extension of narrowband telephone speech can extend the audio bandwidth either below or above the telephone band, or both. In this paper, only bandwidth extension to frequencies above the telephone band is considered. The terms *lowband* and *narrowband* are used to refer to the transmitted frequency band below 4 kHz, whereas *highband* stands for the frequency range 4–8 kHz.

Most of the published speech bandwidth extension methods are based on the source-filter model of speech production. They generate an excitation signal that is modified with a filter simulating the spectral shaping characteristics of the vocal tract. Both the excitation signal and the vocal tract filter are estimated using only the information contained in the narrowband signal. An artificially generated highband signal is then combined with the original narrowband signal to form the wideband signal with extended bandwidth.

Commonly used techniques for generating a wideband excitation signal include spectral folding, spectral translation, and nonlinear processing of an excitation derived from the narrowband signal [7]. Alternatively, sinusoidal synthesis [6] or modulated noise [17] can be used. The vocal tract filter is often realized as an all-pole filter that is parametrized using, e.g., line-spectral frequencies or cepstral coefficients [6]. The parameters of the highband shaping filter can be estimated from narrowband features using, e.g., codebooks [6], Gaussian mixture models (GMM) [9], hidden Markov models [8], or neural networks [6, 10]. The highband can also be constructed from subband signals that are weighted and processed appropriately [9].

An ABE technique without explicit use of the source-filter model was presented in [11]. This method was based on spectral folding of the lowband signal to the highband, classification of speech frames into phonetically motivated classes, and spectral shaping of the highband in the frequency domain with smooth spline curves. Listening tests were arranged to verify the performance of this method with different languages [15]. For a couple of years, the technique has been on the market in several mobile phone models of Nokia [12]. This method is used as a reference algorithm in this paper and it is referred to as Ref-ABE.

The method described in this paper was developed in an attempt to improve the speech quality of the Ref-ABE method. The following three aspects were considered: Firstly, Ref-ABE sometimes fails to reproduce the highband content of sibilant sounds. Using a modified feature set together with a more sophisticated highband estimation technique in the new method was expected to produce more consistent sibilants. Secondly, the overall timbre of Ref-ABE has sometimes been found to be too crisp and to have an unnatural character. These issues were expected to be reduced by a more accurate control of the highband spectral shape, which could be achieved by using a spectrally flat highband excitation instead of modifying the folded lowband spectrum directly. Finally, the new method was chosen to be implemented using only time-domain processing in the signal path, which may be beneficial for practical real-time implementation on some platforms where fixed-point FFT routines may cause audible distortion.

2. METHOD

The proposed bandwidth extension method extends the audio bandwidth of narrowband speech by generating spectral

 ISSN 2076-1465

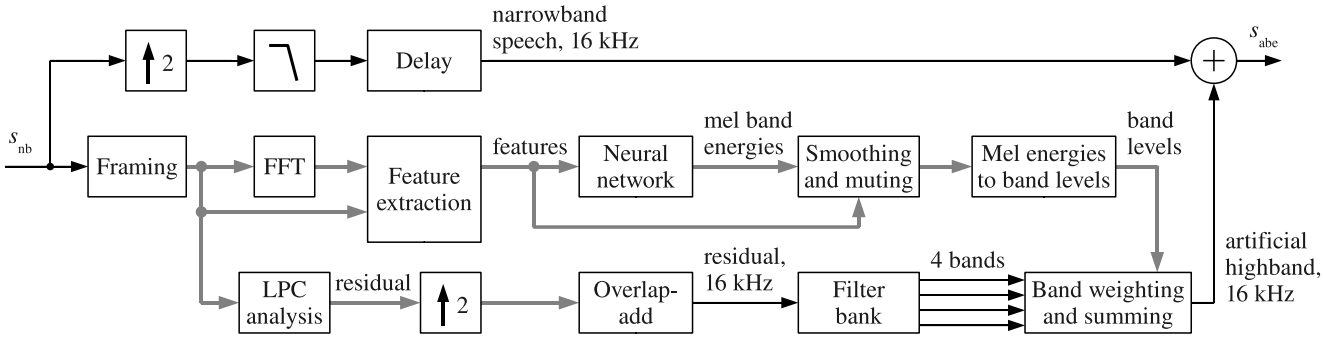

Figure 1: Block diagram of the bandwidth extension method. The narrowband input and the bandwidth-extended output are denoted by s_{nb} and s_{abe}, respectively. Gray arrows indicate frame-based processing, whereas black arrows show signal paths where sample-by-sample processing in the time domain can be utilized. The filter bank is depicted in Figure 2.

content in the frequency range 4–8 kHz. A block diagram of the method is shown in Figure 1.

2.1 Feature extraction

The input signal to the system, s_{nb}, has a sampling frequency of 8 kHz. For analysis, the input signal is divided into frames of 12 ms with the hop size of 10 ms. Experiments with frame sizes of about 10 ms and 20 ms showed negligible difference in quality, and a shorter frame size was chosen due to a smaller framing delay.

A set of time-domain and frequency-domain features is extracted from each frame. The features were selected with the goal of differentiating between various speech sounds that call for different spectral envelope in the highband. The following features were found to constitute an appropriate set with reasonable size:

Energies of five subbands: The frequency range 250–3500 Hz is divided into five subbands with equal widths on the mel scale. Subband energies are calculated from the power spectrum and converted to the decibel scale.

Centroid of the lowband power spectrum: The center of gravity is computed from the lowband power spectrum, and the resulting centroid frequency is squared to improve the performance of the feature. A spectral centroid feature gives higher values for unvoiced speech than for voiced speech [7].

Gradient index: The gradient index is defined as the sum of the signal gradient magnitudes at each change of signal direction [8]. This feature also differentiates between voiced and unvoiced speech segments [7, 15]. It has been found to vary depending on speaker, noise, and speech coding method [11]. Therefore, the feature is normalized by an adaptive long-term estimate of the range of the gradient index values.

Energy ratio: The signal energy of the current frame is divided by that of the previous frame, and the ratio is converted to the decibel scale. This feature has been found to be useful for the differentiation of stop consonants from other unvoiced speech sounds [10, 15].

Spectral flatness: Spectral flatness is the ratio between the geometric and the arithmetic mean of the power spectrum. It is computed from the range 0.3–3.4 kHz. This feature indicates the tonality of the signal [7].

Voice activity detector: The voice activity detector (VAD) defined for the AMR coder [1] (option 2) is used. This allows effective attenuation of the highband when no speech is present. To avoid abrupt changes, the binary VAD output is smoothed to change gradually from 1 to 0 when the end of voice activity is indicated.

2.2 Estimation of the highband spectral envelope

The features are fed to a neural network that estimates the highband spectral shape from the lowband features. The neural network outputs represent signal energy levels of four subbands within the highband. Four subbands were experimentally found to provide adequate frequency resolution. The center frequencies (4595 Hz, 5278 Hz, 6063 Hz, and 6964 Hz) and bandwidths of the subbands were adopted from the commonly used mel filter bank [14].

The neural network was constructed using the method called neuroevolution of augmenting topologies (NEAT) [16]. This method starts from a minimal neural network topology and incrementally improves the network performance not only by modifying the network weights but also by adding new nodes and connections to the network structure. The resulting network is not restricted to any predefined topology but the structure is grown incrementally during the learning process. The NEAT network used in this study had 29 nodes and 164 connections.

The training data consisted of 17 minutes of spoken American English and Finnish from the NTT speech database [13]. The speech signals were first high-pass filtered with the MSIN filter [5], which simulates the input characteristics of a mobile station. To extract narrowband features, the signals were downsampled to 8 kHz sampling rate and processed with the AMR speech codec. The corresponding highband parameters were extracted from unprocessed wideband speech.

Rapid changes in the highband are limited by smoothing the outputs of the neural network with a recursive filter when the output values are rising from the previous frame. To reduce noise during pauses, the highband is attenuated when the power in the telephone band is close to an adaptive noise level estimate.

2.3 Highband synthesis

The synthesis part of the method produces an artificial highband signal according to the estimated subband energy levels. Only time-domain processing is used in the signal path. A residual signal given by linear predictive coding (LPC)

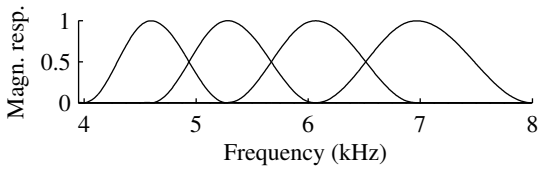

Figure 2: Filter bank with four subbands in the highband.

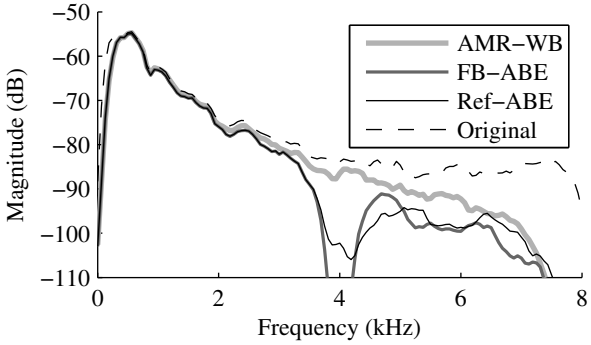

Figure 3: Comparison of long-term average spectra.

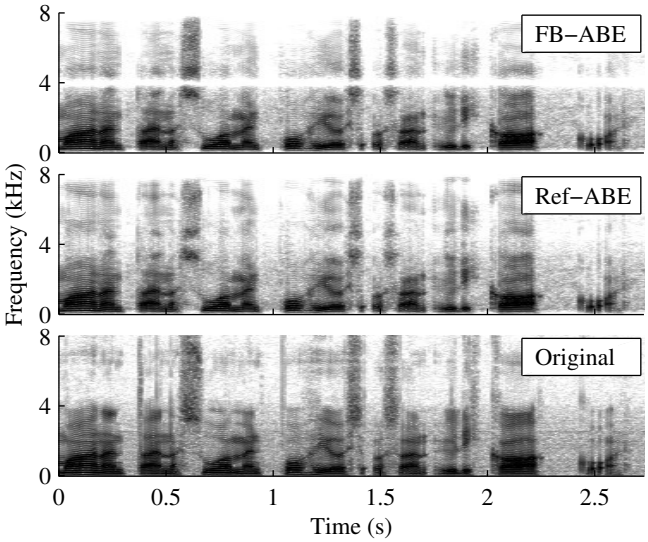

Figure 4: Comparison of spectrograms.

is first computed from the narrowband speech frame using the autocorrelation method and prediction order of 10. The residual is modified with a formant filtering technique similar to the short-term postfiltering used in some speech coders for noise reduction [2]. This method amplifies formant frequencies and attenuates spectral valleys, and thus restores a weak formant structure to the residual in order to attenuate possible noise at spectral valleys. This technique was found to slightly reduce perceived noise in the highband. Each modified residual frame is scaled in amplitude to constant average power, and successive frames are combined with overlap-add. A wideband excitation is then obtained from this modified residual by spectral folding, i.e., adding zeros between the samples. Compared to the direct shaping of folded lowband signal used in the Ref-ABE method [11], the use of the residual is supposed to improve the naturalness of the resulting extension band signal because strong formant trajectories are not mirrored from the lowband to the highband.

The highband of the excitation signal is divided into four frequency bands using a filter bank of linear-phase 128-tap FIR filters. The filters are designed such that the passbands of adjacent filters overlap and their magnitude responses sum to unity. The center frequencies of the filters correspond to those of the mel filter bank for the highband. Figure 2 shows the magnitude responses of the four bandpass filters.

A mapping from the estimated subband energies to the corresponding gain coefficients of passband signals is computed using an iterative technique that assumes a spectrally white excitation signal. An initial estimate of gain values is obtained using the assumption that the passband signals do not affect adjacent mel bands, and the gain coefficients are then corrected in each iteration by the ratio of the target mel spectrum to the currently resulting mel spectrum. Target values for the subband gain coefficients are computed for each speech frame, and the actual subband weights are changed smoothly at frame boundaries. The passband signals are weighted and summed to produce a highband signal.

The lowband signal is upsampled and lowpass filtered. The output of the system, s_{abe}, is obtained by adding the artificially generated highband signal to the appropriately delayed lowband signal.

As an alternative to weighting passband signals, a single FIR filter could be computed for the highband as a weighted sum of the bandpass filters. A similar filter bank equalizer technique was used in [3]. Decimation of intermediate signals could also be used to reduce computation.

The delay caused by the algorithm is due to framing, overlap-add, and filtering. Framing and overlap-add cause a delay equal to the frame length of 12 ms and the filter bank delays the signal by additional 4 ms.

The proposed algorithm is referred to as the filter bank based ABE (FB-ABE) in the rest of the paper.

3. EVALUATION

This section presents spectral comparisons and subjective evaluations that are based on 10 speech excerpts spoken in Finnish by ten different talkers (five females and five males). Each speech excerpt is between 2 and 7 seconds of length and contains one spoken sentence. The speech signals have been recorded with high-quality equipment in an anechoic chamber and stored digitally using a sampling rate of 44.1 kHz.

Test samples simulating realistic cellular telephone speech were generated from the high-quality recordings. The signals were downsampled to 16-kHz sampling frequency and high-pass filtered with the MSIN filter, which approximates the input response of a mobile station. Speech level was normalized to -26 dBov, and MSIN-filtered office noise was added at a signal-to-noise ratio of 35 dB. The preprocessed signals were downsampled to 8-kHz sampling rate and processed twice through the AMR codec at 12.2 kbps, thus simulating the suboptimal case in which encoding and decoding are performed twice in the transmission path. The coded speech samples were then processed with Ref-ABE [11] and the proposed FB-ABE method. For reference, wideband telephone speech was also simulated by passing the preprocessed signals twice through the AMR-WB codec using the bit rate of 12.65 kbps. Finally, the frequency response of a wideband mobile terminal was approximated by filtering all samples with a band-pass filter having cutoff frequencies at about 270 Hz and 7300 Hz.

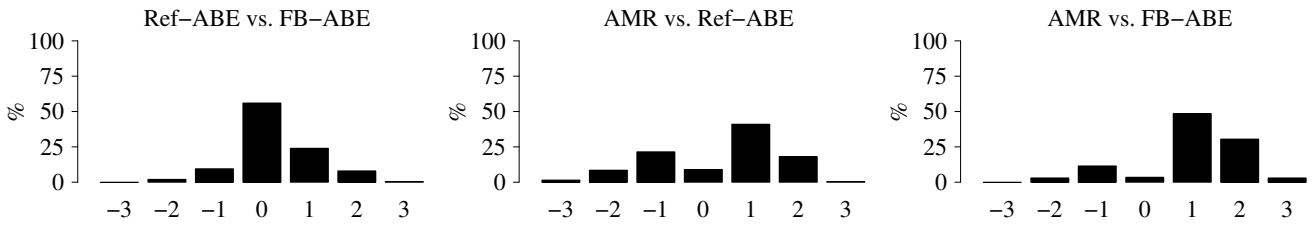

Figure 5: Distributions of listener ratings in pairwise comparisons between Ref-ABE and FB-ABE, AMR and Ref-ABE, and AMR and FB-ABE. The bars indicate relative frequencies of the scores from much worse (−3) to much better (3).

3.1 Comparison of spectra

Long-term average spectra of the signals processed with AMR-WB and the two bandwidth extension systems are shown in Figure 3. The average spectrum of the original speech signals after high-pass filtering, scaling, and noise addition is also shown for comparison. The highband spectra of Ref-ABE and FB-ABE have approximately similar shape and level on average. Both bandwidth extension methods produce a lower highband level than AMR-WB, which also attenuates the highband compared with the original wideband signal. Both ABE methods also show a gap in the spectrum at 4 kHz. Such a narrow stopband has been experimentally found to have only a minor perceptual effect as described previously in [8, 15].

Figure 4 shows spectrograms of a speech segment processed with FB-ABE and Ref-ABE. The spectrogram of the original speech segment after high-pass filtering, scaling, and noise addition is also shown for reference. Folded copies of lowband formants in the highband are less pronounced in FB-ABE than in Ref-ABE. FB-ABE also regenerates the spectrum of sibilant sounds more accurately than Ref-ABE.

3.2 Listening test

A formal listening test was arranged to compare the proposed FB-ABE method with the existing Ref-ABE method and with narrowband (AMR) and wideband (AMR-WB) references. The listening test procedure was similar to the Comparison Category Rating (CCR) test described in [4]. The test consisted of pairwise comparisons between the processing types. One sentence, processed in two different ways, was presented to the listener in each test case. The listener was asked to evaluate the quality of the second sample in comparison with the quality of the first sample. Responses were given using the seven-point comparison mean opinion score (CMOS) scale ranging from much worse (−3) to much better (3), zero indicating about the same quality. Listeners were allowed to repeat each sample pair with no limitations before answering. The test was arranged separately for each listener in a quiet room using a graphical user interface on a computer screen, and test samples were played to both ears through Sennheiser 580 headphones. Each listener had a short practice session before starting the actual test. The subjects were allowed to adjust the volume setting to a suitable level during the practice session.

Twenty listeners (5 females and 15 males) between 20 and 33 years of age participated in the test. The listeners were native speakers of Finnish and none of them had any known hearing defects.

For each of the 10 test sentences, all six combinations of the four processing types were presented to each listener. The test included an equal number of comparisons in both presentation orders for each pair of processing types. For the evaluation of listener consistency, all comparisons of two of the test sentences were presented in both presentation orders, and 12 null pairs of identical samples were also included in the test. Altogether, the test comprised 84 comparisons. The order of the test items was randomized separately for each listener using some balancing constraints on the order.

No listeners were excluded from the data based on their responses. Duplicated evaluations of the same sample pairs as well as null pairs were excluded from the data before performing the following analysis.

The distributions of ratings in comparisons between each pair of processing types was collected. For brevity, Figure 5 presents the score distributions obtained for the three most relevant comparisons. The bars show the relative frequency of each score given in the comparisons between the two procesing types. Bars on the positive side indicate preference for the latter of the processing types shown in the illustration title. The order of presentation was normalized such that the scores were negated in the cases where the processing types were actually presented in the opposite order.

In comparisons between Ref-ABE and FB-ABE, the quality was assessed to be about the same in most of the cases. However, the distribution of ratings shows some preference towards the FB-ABE method. The distributions of comparisons involving the narrowband reference (AMR) and either of the ABE methods show two peaks, one indicating preference for the ABE-processed sample and the other for the narrowband reference. The number of cases where the narrowband sample was preferred is, however, considerably smaller in the comparisons between AMR and FB-ABE than in those between AMR and Ref-ABE.

Another visualization of the listener ratings is shown in Figure 6. The mean score for each processing type was calculated from all the comparisons in which the processing was involved. Again, the presentation order of the sample pairs was normalized and the scores negated as necessary. This method yields the order of superiority and distances between the processing types, but the mean scores cannot be interpreted using the CMOS scale.

The figure shows that both ABE methods were rated significantly better than the narrowband reference, while the quality of the wideband reference processing (AMR-WB) was considered much better than that of either of the ABE methods. An important result is that FB-ABE was found to have significantly higher quality than Ref-ABE.

3.3 Observations in informal listening evaluations

Subparts of the FB-ABE algorithm were tested separately, e.g., using correct highband mel spectra together with high-

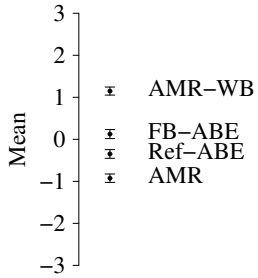

Figure 6: The order of preference of the processing types in the listening test. Mean scores and 95 % confidence intervals are shown.

band excitation generated from the lowband residual signal. Such experiments verified that the reconstruction of the highband mel spectrum produces good perceptual quality, whereas the estimation of the highband mel spectrum is the most critical part and the source of most of the quality degradation.

Informal listening comparisons between FB-ABE and Ref-ABE using a wider set of test samples than in the formal test indicated that FB-ABE improves the overall timbre of bandwidth-extended speech, and the quality of sibilants is also improved. Presumably, these quality enhancements account for the results of the formal listening tests showing preference for the proposed FB-ABE method. On the other hand, FB-ABE was found to be more sensitive to breathing sounds, which cause some additional noise in the highband.

4. CONCLUSIONS

A new FB-ABE method for the bandwidth extension of telephone speech was introduced and a subjective evaluation of the method was presented. The method generates artificially spectral content in the band 4–8 kHz. It utilizes a neural network to control the weights of passband excitation signals that constitute the extension band signal. Only time-domain processing is used in the signal path, which may be beneficial for real-time implementation on some platforms. The small number of subbands in the highband also permits implementation with moderate computational cost.

The FB-ABE method was developed with the goal of achieving higher speech quality than what is provided by the Ref-ABE method, which is used in some Nokia mobile phone models and has earlier been shown to improve both quality and intelligibility of narrowband speech [15, 12]. Listening test results presented in this paper indicate that significant quality improvement was indeed achieved, but FB-ABE also involves substantially more computation than Ref-ABE. Further work is planned on reducing highband noise that is sometimes generated by FB-ABE during breathing sounds. More extensive testing is also needed to evaluate the method in different languages and background noise conditions.

5. ACKNOWLEDGEMENTS

The work of H. Pulakka is funded by the Finnish Graduate School in Electronics, Telecommunications and Automation (GETA) and by Nokia Devices. Pulakka is supported by the Foundation of Nokia Corporation, Tekniikan edistämissäätiö (TES), and the Finnish Foundation for Economic and Technology Sciences (KAUTE, Kaartokulma's fund).

REFERENCES

[1] 3rd Generation Partnership Project (3GPP). *Adaptive Multi-Rate (AMR) Speech Codec, Voice Activity Detector (VAD), 3GPP TS 26.094*, 2007. Version 7.0.0.

[2] J.-H. Chen and A. Gersho. Adaptive postfiltering for quality enhancement of coded speech. *IEEE Trans. Speech Audio Process.*, 3(1):59–71, 1995.

[3] B. Geiser, P. Jax, P. Vary, H. Taddei, S. Schandl, M. Gartner, C. Guillaumé, and S. Ragot. Bandwidth extension for hierarchical speech and audio coding in ITU-T Rec. G.729.1. *IEEE Trans. Audio, Speech, Language Process.*, 15(8):2496–2509, 2007.

[4] International Telecommunication Union. *ITU-T Recommendation P.800, Methods for subjective determination of transmission quality*, 1996.

[5] International Telecommunication Union. *ITU-T Recommendation G.191, Software tools for speech and audio coding standardization*, 2005.

[6] B. Iser and G. Schmidt. Bandwidth extension of telephony speech. *EURASIP Newslett.*, 16(2):2–24, 2005.

[7] P. Jax. *Enhancement of Bandlimited Speech Signals: Angorithms and Theoretical Bounds*. PhD thesis, Rheinisch-Westfälische Technische Hochschule Aachen, 2002.

[8] P. Jax and P. Vary. On artificial bandwidth extension of telephone speech. *Signal Process.*, 83(8):1707–1719, 2003.

[9] K.-T. Kim, M.-K. Lee, and H.-G. Kang. Speech bandwidth extension using temporal envelope modeling. *IEEE Signal Process. Lett.*, 15:429–432, 2008.

[10] J. Kontio, L. Laaksonen, and P. Alku. Neural network-based artificial bandwidth expansion of speech. *IEEE Trans. Audio, Speech, Language Process.*, 15(3):873–881, 2007.

[11] L. Laaksonen, J. Kontio, and P. Alku. Artificial bandwidth expansion method to improve intelligibility and quality of AMR-coded narrowband speech. In *Proc. ICASSP*, pages 809–812, 2005.

[12] L. Laaksonen, H. Pulakka, V. Myllylä, and P. Alku. Development, evaluation and implementation of an artificial bandwidth extension method of telephone speech in mobile terminal. *IEEE Trans. Consum. Electron.*, 55(2):780–787, 2009.

[13] NTT Advanced Technology Corporation. *Multi-Lingual Speech Database for Telephonometry*, 1994.

[14] J. W. Picone. Signal modeling techniques in speech recognition. *Proc. IEEE*, 81(9):1215–1247, 1993.

[15] H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen, and P. Alku. Evaluation of an artificial speech bandwidth extension method in three languages. *IEEE Trans. Audio, Speech, Language Process.*, 16(6):1124–1137, 2008.

[16] K. O. Stanley and R. Miikkulainen. Evolving neural networks through augmenting topologies. *Evol. Computation*, 10(2):99–127, 2002.

[17] T. Unno and A. McCree. A robust narrowband to wideband extension system featuring enhanced codebook mapping. In *Proc. ICASSP*, pages 805–808, 2005.