ARTIFICIAL BANDWIDTH EXPANSION METHOD TO IMPROVE INTELLIGIBILITY AND QUALITY OF AMR-CODED NARROWBAND SPEECH

Laura Laaksonen

Nokia Research Center
Multimedia Technologies Laboratory
P.O.Box 407, 00045 Nokia Group, FINLAND
e-mail: laura.laaksonen@nokia.com

Juho Kontio and Paavo Alku

Helsinki University of Technology
Lab. of Acoustics and Audio Signal Processing
P.O.Box 3000, 02015 HUT, FINLAND
e-mail: jkontio@acoustics.hut.fi, paavo.alku@hut.fi

ABSTRACT

Speech quality suffers from limited bandwidth of cellular telephone systems, making it sound muffled. In addition, intelligibility is degraded due to missing higher frequency components. The proposed enhancement system is designed to improve both intelligibility and quality of narrowband speech by expanding the bandwidth and creating new spectral components to high frequencies in the receiving end of the transmission link. The algorithm can be used together with conventional narrowband speech codecs and it is designed to be robust in different noise conditions. In addition, the computational load of the algorithm is reasonable.

1. INTRODUCTION

Speech is transmitted in communication networks mostly in narrowband by using audio bandwidth (0.3-3.4 kHz) and sampling frequency (8 kHz) originating from conventional PCM. Because of the limited frequency range the quality of this, so called telephone speech, is not natural. Both pleasantness and intelligibility suffer from limited bandwidth. Therefore, wideband speech transmission (up to 7 kHz) will become more prevalent in the future and wideband speech codecs, such as AMR-WB for GSM and UMTS [1], have been standardized. However, it will take time before all the terminals and networks support wideband transmission. Meanwhile another approach towards wideband transmission is to *artificially* add new upper spectral components to speech in the receiving end of the transmission link without any additional transmitted information. In this *artificial bandwidth expansion* (ABE) approach, the speech signal is transmitted through the network using conventional narrowband codecs and the receiver expands the bandwidth using an ABE method. This way it is possible to obtain speech of better quality and intelligibility with existing communication systems and the quality gap between narrowband and wideband transmission is narrowed.

During the last years, many approaches of artificial bandwidth expansion have been proposed. The quality of artificially created high band depends on how well the prediction of the fine structure and the envelope of the high band spectrum has been succeeded. In spectral folding, the fine structure is generated by up-sampling the original narrowband signal, which means that the frequency components of the high band are obtained as a mirror image of the narrowband spectrum [11]. Spectral translation, i.e. creating a copy of the narrowband spectrum to produce the high band, is another simple approach [6]. Other techniques to create new frequency components to the high band are by non-linear processing [6] or by using sinusoidal signals [2].

The main focus in artificial bandwidth expansion algorithms has been on envelope prediction methods. Since no extra information is transmitted, the envelope prediction is a challenging task. Early methods used a fixed filter to modify the high band [11]. Later, more sophisticated methods have been proposed. These methods include, for example, codebook mapping, where a high band envelope is selected from a pre-designed codebook [2],[3]. The best wideband envelope is chosen by comparing the original narrowband envelope to the wideband entries in the codebook. Another approach, closely related to codebook mapping, is to use HMM (Hidden Markov Models) [6]. In an HMM, each state corresponds to a specific speech sound, just like a codeword in a codebook. GMMs (Gaussian Mixture Model) have also been used for predicting wideband envelopes [8]. Narrowband feature vectors are mapped to wideband envelopes using a GMM that has been taught by a training set. Recently, also methods based on the neural networks have been proposed [9]. They can be used to create the fine structure and/or the envelope of the high band.

The ABE method described in this paper is based on spectral folding. The envelope prediction technique is a

novel approach that is based on cubic splines. The proposed enhancement system was designed to improve both intelligibility and quality of coded narrowband speech. In addition, the algorithm is robust in different noise conditions and it can be implemented with a reasonable computational cost. Differently from previous methods of artificial bandwidth expansion using spectral folding, this algorithm classifies the speech sounds into three different categories and adjusts the modification of the high band envelope correspondingly. This makes it possible to focus on certain types of envelopes at a time.

2. DESCRIPTION OF THE ALGORITHM

The block diagram of the proposed algorithm is presented in Figure 1. The input signal, s_{nb}, is a narrowband signal sampled with sampling frequency of 8 kHz. First, the input signal is divided into frames and up-sampled by 1:2. As a result the sampling frequency is doubled to 16 kHz and the aliased frequency components are created to the high band, i.e. to the frequency range of 4-8 kHz. These aliased frequency components are mirror images of the spectral samples in the lower band from 0 Hz to 4 kHz.

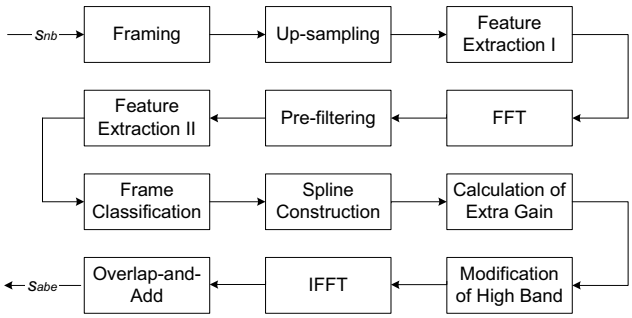

Figure 1: The block diagram of the proposed algorithm. The input, s_{nb}, denotes the original narrowband signal and the output, s_{abe}, the expanded wideband signal.

A set of time domain features is calculated from the original narrowband signal. These features are used in the frame classification that is discussed later. The modification of the mirror images of the high band is made in the frequency domain with the help of FFT.

A disadvantage of spectral folding is the violation of the harmonic structure in the high band because upper harmonics are mirror images of the lower one and therefore they do not (necessarily) locate at multiple integers of the fundamental frequency. Even though it is known that the envelope of the high band is more important perceptually than the fine structure, it was noticed that some distortion appeared when the upper harmonics were strong. This problem was solved by pre-filtering the amplitude spectrum in the frequency domain after FFT. A simple filter was designed to smoothen the mirror images of the harmonic components in the frequency band of 5.5-8 kHz. The filter is of form $H(z)=-0.7/(1+0.2z^{-1})$.

A frequency domain feature, narrowband slope, is calculated from the pre-filtered mirror image spectrum in the high band to complete the feature vector. The classification of the frame into one of the three speech sound categories is made based on the feature vector.

The modification function for the high band is based on cubic splines. They are smooth and continuous curves and can be controlled with only a few parameters. The original lower band is left untouched. The spline curve is constructed around five control points that are at the intervals of 1 kHz, i.e. at 4 kHz, 5 kHz, 6 kHz, 7 kHz and 8 kHz. Magnitudes for control points are calculated using a speech sound class specific equation. After the control points have been defined, the modification curve is constructed using cubic spline interpolation with not-a-knot end conditions.

The level of the spline curve is further adjusted with an extra gain that is both near-end and far-end noise dependent to improve the performance in noisy conditions. The amplitude spectrum of the high band is multiplied by the modification curve and finally the wideband speech frame is obtained by inverse Fourier-transforming the modified spectrum and adding the frames together with an overlap-and-add technique.

2.1. Frame Classification

Frames are classified into three categories; voiced sounds, sibilants and stop-consonants. The frame classification procedure is based on a set of features calculated from the original narrowband signal. These features include:

- Gradient index, which is defined as a measure of the sum of the magnitudes of the gradient of the speech signal at each change of direction [6].
- Gradient count, which is a feature describing how long the level of gradient indices has exceeded a predefined level.
- Energy ratio, which is the energy of the current frame divided by the energy of the previous frame.
- Energy quotient, which is a ratio between a short term frame energy and a long term frame energy.
- Narrowband slope (n_{nb}), which is a slope of the narrowband amplitude spectrum. The slope is estimated from amplitude spectrum between frequencies of 0.3 and 3.0 kHz.

Each feature of the feature vector is compared to its threshold value and the classification is made based on the comparisons. All other threshold values are fixed except for the threshold for the gradient index, which is made adaptive in order to improve the accuracy of sibilant

detections. The threshold follows the long-term level of gradient index, since it varies depending on the speaker, background noise and the speech coding method.

2.2. Modification of High Band

Adaptive cubic splines with five control points are used to shape the high band in the frequency domain. The control points consist of a fixed base part and an adaptive part, which uses narrowband slope parameter (n_{nb}) to modify the control point magnitude. For control point k, the final magnitude is defined as:

$$C_k = b_k + a_k \cdot n_{nb}, \quad 1 \le n \le 5 \qquad (1)$$

where b_n and a_n are predefined control point constants for control point k.

For each frame category, a separate set of control point constants is used. The constants are optimized offline using a genetic algorithm (GA) [4] based search. The GA searches for control point constant values that minimize the spectral MSE of the frame expansions in a teaching sample set. An example of a spline curve and its effect on the amplitude spectrum is presented in Figure 2.

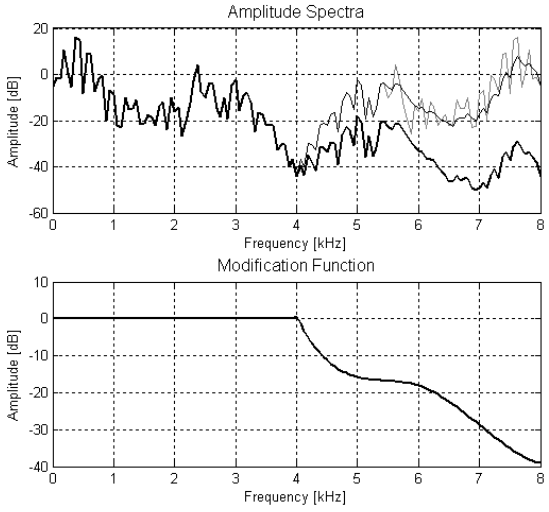

Figure 2: Amplitude spectrum of */i/* spoken by a female speaker (top) and the corresponding modification function (bottom). Gray curve is the folded spectrum. Thin black curve is the spectrum after pre-filtering and bold black curve the modified spectrum.

The classification of speech sounds becomes more difficult as the far-end background noise level increases. In addition, the bandwidth expansion has to be performed more carefully because too aggressive expansion can easily result in an increased noise level. On the other hand, the noise at the listener masks possible artifacts and the attenuation does not have to be so large. The algorithm is made background noise dependent in order to guarantee the best performance in every noise situation. The level of the designed spline curve is adjusted with noise dependent gain that is constant over the entire high band.

Background noise levels are estimated using a method called minimum statistics [7] and signal-to-noise ratio (SNR) estimates are then derived from them. If the SNR estimate between far-end speech and far-end noise is less than 30 dB, an extra attenuation gain is applied to the spline function. The attenuation increases up to 25 dB as the SNR decreases. If the SNR estimate between far-end speech and near-end noise is less than 35 dB, an extra amplification gain is applied to the spline function. The amplification increases up to 10 dB as the SNR decreases. An example case is presented in Figure 3. An extra attenuation gain of 8 dB is applied to the high band.

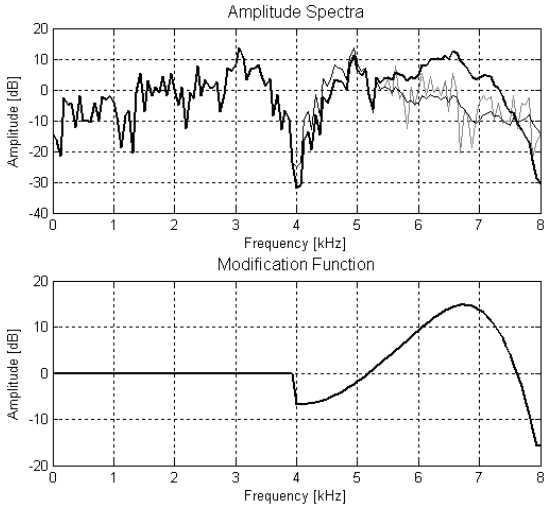

Figure 3: Amplitude spectrum of */s/* spoken by a female speaker (top) and the corresponding modification function (bottom). Gray curve is the folded spectrum. Thin black curve is the spectrum after pre-filtering and bold black curve the modified spectrum.

3. RESULTS

Subjective listening tests were arranged in order to measure the intelligibility improvement and the overall quality of the algorithm. The intelligibility test was a speech reception threshold (SRT) in noise test described in [10]. The speech reception thresholds were measured in three different noises; car noise, babble noise and narrowband SRT noise that has an average speech spectrum. In all cases ABE samples had lower SRT-level and thus were more intelligible than corresponding original narrowband sounds. In SRT noise, ABE samples were even more intelligible than wideband samples.

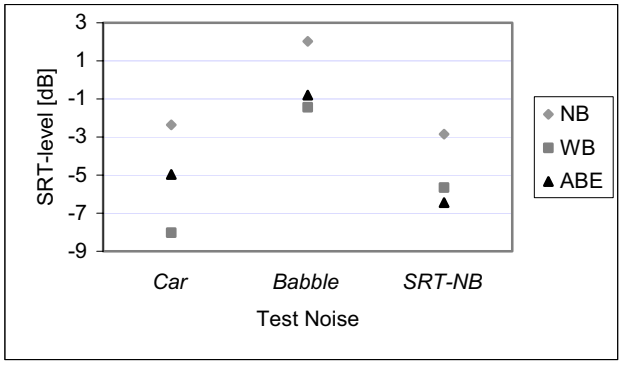

Figure 4: Test results from speech reception threshold (SRT) in noise test. Thresholds are measured for narrowband (NB), wideband (WB) and artificial bandwidth expanded (ABE) speech in three different noise conditions (car, babble and SRT noise).

The quality test was carried out following the ITU-T P.830 Recommendation [5], which describes subjective assessment of codecs. The test results are shown in Figure 5. There are three processings, AMR-WB coded (with bit rate of 12.65 kbit/s) wideband speech, AMR coded (with bit rate of 12.2 kbit/s) narrowband speech and ABE (applied to AMR coded narrowband speech) processed speech. 24 listeners participated in the test. The results indicate that ABE processed samples tend to have better quality than narrowband speech. When comparing to other bandwidth expansion algorithms it should be kept in mind that this algorithm does not require large codebooks and the computational requirements are relatively small.

The subjective tests also revealed that the algorithm is somewhat speaker dependent so that the quality is improved more for some speakers than for others. The highest improvement score, 0.5 MOS scores, was obtained for a male speaker.

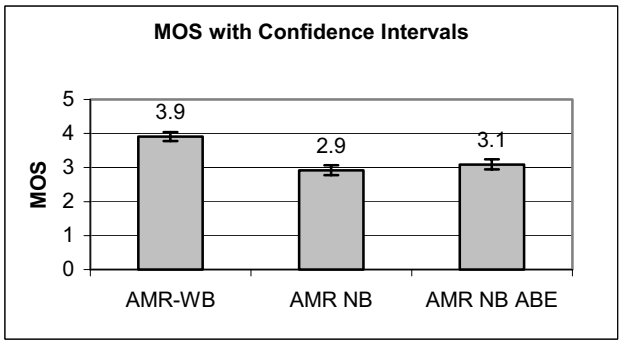

Figure 5: Test results from the quality test. MOS scores and 95% confidence intervals are presented for AMR-WB coded (bit rate of 12.65 kbit/s) wideband speech (AMR-WB), AMR coded (bit rate of 12.2 kbit/s) narrowband speech (AMR NB) and ABE (applied to AMR coded narrowband speech) processed speech (AMR NB ABE).

4. CONCLUSIONS

The proposed ABE algorithm doubles the sampling frequency of conventional narrowband speech and adds new spectral components to the high band. The algorithm was designed to meet the requirements set by real telecommunication systems. Firstly, the algorithm was to be designed for coded speech, secondly it should be robust in different noise situations, and thirdly the computational requirements were to be reasonable.

Subjective listening tests proved that the algorithm improves the intelligibility and quality of coded narrowband speech.

5. REFERENCES

[1] 3GPP TS 26.171, AMR wideband speech codec; general description, 2001.

[2] C.-F. Chan and W.-K. Hui, "Quality Enhancement of Narrowband CELP-coded Speech via Wideband Harmonic Re-synthesis", IEEE Proc. of Int. Conf. on Acoustics, Speech and Signal Proc., pp.1187-1190, 1997.

[3] J. Epps and W.H. Holmes, "A New Technique for Wideband Enhancement of Coded Narrowband Speech", Proc. of the IEEE Workshop on Speech Coding, pp.174-176, 1999.

[4] J.H. Holland, "Adaptation in Natural and Artificial Systems", University of Michigan Press, 1975.

[5] ITU-T P.830, Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs, 1996.

[6] P. Jax, "Enhancements of Bandlimited Speech Signals", Ph.D thesis, RWTH Aachen, vol. 15 of P. Vary (ed.), Aachen Beiträge zu digitalen Nachrichtensystemen, 2002.

[7] R. Martin, "Spectral Subtraction Based on Minimum Statistics," Proc. Eur. Signal Proc. Conf., pp. 1182-1185, 1994

[8] K.-Y. Park and H.S. Kim, "Narrowband to Wideband Conversion of Speech Using GMM Based Transformation", IEEE Proc. of the Int. Conf. on Acoustics, Speech and Signal Proc, pp. 1843-1846, 2000.

[9] A. Uncini, F. Gobbi and F. Piazza, "Frequency Recovery of Narrow-band Speech Using Adaptive Spline Neural Networks", IEEE Proc. of the Int. Conf. on Acoustics, Speech and Signal Proc, pp. 997-1000, 1999.

[10] M. Vainio, A. Suni, H. Järveläinen, J. Järvikivi, V.-V. Mattila, "Development of a Speech Intelligibility Test for Mobile Communications Based on Measuring Speech Reception Thresholds in Noise" (Submitted to JASA)

[11] H. Yasukawa, "Quality Enhancement of Band Limited Speech by Filtering and Multirate Techniques," Proc. Of Int. Conf. On Spoken Language Proc., pp.1607-1610, 1994.