Subjective Voice Quality Evaluation of Artificial Bandwidth Extension: Comparing Different Audio Bandwidths and Speech Codecs

Hannu Pulakka¹, Anssi Rämö², Ville Myllylä³, Henri Toukomaa², Paavo Alku³
¹Lumia Audio Technology, Microsoft, Tampere, Finland
²Nokia Research Center, Tampere, Finland
³Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

hannu.pulakka@microsoft.com

Abstract
Artificial bandwidth extension (ABE) methods have been developed to improve the quality and intelligibility of telephone speech. In many previous studies, however, the evaluation of ABE has not fully reflected the use of ABE in mobile communication (e.g., evaluation with clean speech without coding). In this study, the subjective quality of ABE was evaluated with absolute category rating (ACR) tests involving both clean and noisy speech, two cutoff frequencies of highpass filtering, and input encoded at different bit rates. Three ABE methods were evaluated, two for narrowband-to-wideband extension and one for wideband-to-superwideband extension. Several speech codecs with different audio bandwidths were included in the tests. Narrowband-to-wideband ABE methods were found to significantly improve the speech quality when no background noise was present, and the mean quality scores were slightly but not significantly increased for noisy speech. Wideband-to-superwideband ABE also showed significant improvement in certain conditions with no background noise. ABE did not cause significant decrease of the mean scores in any of the tests.

Index Terms: artificial bandwidth extension, subjective evaluation, listening test, speech coding

1. Introduction
Speech transmission in communication networks is still commonly limited to narrowband speech with an audio band constrained below 4 kHz. The adaptive multi-rate (AMR) codec [1] widely used in mobile networks is an example of a narrowband speech codec. Better speech quality and intelligibility can be obtained by transmitting wideband speech with an audio band of 50–7000 Hz. Wideband speech services are currently being deployed in a growing number of mobile networks [2] using the adaptive multi-rate wideband (AMR-WB) codec [3]. However, natural speech contains frequency content beyond the wideband range and the speech quality can be further enhanced using superwideband codecs such as the superwideband mode of the Opus codec [4], which covers frequencies up to 12 kHz, or ITU-T G.722.1 Annex C [5] or ITU-T G.718 Annex B [6], which transmit frequencies up to 14 kHz.

Artificial bandwidth extension (ABE) methods (e.g., [7, 8, 9]) have been developed to extend the audio band of narrowband speech to the wideband frequency range (NB-to-WB) at the receiving end without additional transmitted information. The goal of ABE is to improve the quality and intelligibility of narrowband speech. Furthermore, ABE reduces the difference between narrowband and wideband speech perceived between and within telephone calls [10]. ABE techniques have also been proposed to extend the bandwidth of wideband speech to the superwideband range (WB-to-SWB) [11, 12, 13].

The subjective quality of ABE output can be evaluated with listening test methods defined in [14], which are typically used for the quality characterization of speech codecs. For example, ABE has been evaluated with absolute category rating (ACR) tests in [10, 15] and with comparison category rating (CCR) tests in [16, 17, 9]. The MUSHRA test method described in [18] has also been used (e.g., [19]). Furthermore, conversational evaluations of ABE have been organized [20, 21]. In most of the published evaluations, ABE has been found to improve the speech quality (e.g., [16, 19, 9]), but especially the listening tests reported in [22] and recently in [23] did not show significant improvement over narrowband speech. Intelligibility evaluations have also been arranged (e.g., [10, 24]) showing that ABE can improve the intelligibility of narrowband speech.

ABE methods have often been evaluated with clean speech without speech coding or background noise. However, realistic use of ABE in mobile communication implies that a speech codec is used and downlink noise may be present. ABE evaluations with coded speech have been presented, e.g., in [25, 9], and noise-robust ABE has been considered, e.g., in [26, 22].

This paper presents a subjective evaluation of ABE methods for both clean and noisy speech encoded with different bit rates of the AMR and AMR-WB codecs. Three ABE methods were evaluated: the NB-to-WB ABE method proposed in [9], a new NB-to-WB ABE method based on [9] but employing a different estimation technique, and a similar method for WB-to-SWB extension. The evaluation comprised ACR listening tests similar to those used for codec performance characterization, e.g., in [27, 28]. Several standardized speech codecs with different audio bandwidths were included in the tests, and two highpass filtering cutoff frequencies were also involved.

2. Artificial bandwidth extension methods
This section describes the ABE methods evaluated in this work.

2.1. ABE1: Estimation using a neural network
An ABE method for the extension of narrowband speech (0–4 kHz, 8-kHz sampling) to the wideband frequency range (0–8 kHz, 16-kHz sampling) was proposed in [9]. This method is referred to as ABE1 in this paper. ABE1 uses a neural network to estimate the highband spectrum parameters and a filter bank technique to shape the spectrum. The method was earlier shown to improve the quality of narrowband speech with CCR listening tests in [9] and with conversational tests in [20, 21].
2. ABE2: Estimation using a HMM and linear mapping

A new ABE method was developed with the goal of improving the consistency of output quality for different talkers and reducing artifacts for non-speech sounds such as breathing. A flow diagram of the method is shown in Figure 1. The method is referred to as ABE2 in this work. ABE2 shares the basic structure with ABE1 with the following main differences:

- The synthesis filter bank consists of four subbands with linear spacing in the range 4–8 kHz.
- The feature vector was modified: the number of subbands of the input spectrum was increased to 18. The voice activity detector was removed and a new feature based on the modulation spectrum [29] was added to represent temporal modulation in the input spectrum.
- The neural network was replaced by a hidden Markov model (HMM) and state-specific linear mapping to estimate the highband spectral shape.

The estimation technique is similar to the Gaussian mixture model (GMM) based piecewise linear mapping techniques in [26] and [30], but a HMM is used instead of a GMM. HMM-based ABE techniques have been described, e.g., in [8, 31, 32].

Input features of three successive frames are concatenated to form the feature vector $\mathbf{x}$. The input dimension is reduced using a transformation matrix $\mathbf{L}$ precomputed with linear discriminant analysis (LDA). The resulting vector $\mathbf{z} = \mathbf{Lx}$ is employed by a HMM to compute the probability $p(k|\mathbf{z})$ of each state $k$. An estimate $\hat{y}$ of the subband energy levels in the highband is obtained as a weighted sum of state-specific estimates that are calculated from the input features $\mathbf{x}$ with linear mapping matrices $\mathbf{A}_k$:

$$\hat{y} = \sum_{k=1}^{K} p(k|\mathbf{z}) \mathbf{A}_k [\mathbf{x}^T 1]^T$$

(1)

The HMM, mapping matrices $\mathbf{A}_k$, and LDA matrix $\mathbf{L}$ were trained using 128 minutes of conversational recordings in Finnish with additive noise in part of the training material.

2.3. SWB-ABE: WB-to-SWB extension based on ABE2

Another ABE method was developed for the bandwidth extension of wideband speech (0–8 kHz, 16-kHz sampling) to superwideband speech (0–16 kHz, 32-kHz sampling). This method is referred to as SWB-ABE. The method is based on the same structure as ABE2 with the following major differences:

- The following input features were selected based on mutual information analysis [33] and experiments: gradient index [33], spectral centroid [33], spectral flatness [33], energy quotient [34], differential energy ratio [17], and the input spectrum represented by the energy levels of linearly spaced subbands in the range of 0–7 kHz.
- The excitation is constructed from the linear prediction residual of the input by filtering, modulation, and spectral folding so that the extension band is filled with spectral components of the residual in the range 2–6 kHz. White noise excitation is used for unvoiced speech.
- The synthesis filter bank comprises four linearly spaced subbands in the frequency band 7–15 kHz.
- The extension band is attenuated by 10 dB relative to the level based on training. The attenuation was set experimentally with the aim of reducing the audibility of occasional artifacts and a buzzing character of the extension but maintaining the effect of the extended bandwidth.

3. Subjective evaluation

A subjective listening evaluation was organized to characterize the quality of ABE-processed speech in comparison with narrowband, wideband, and superwideband speech codecs. A similar test setting was used for codec evaluation, e.g., in [27] and [28]. The following conditions were included in the evaluation:

- Direct reference conditions with no speech coding but limited frequency range. Four lowpass cutoff frequencies were evaluated: 4 kHz, 7 kHz, 10 kHz, and 14 kHz.
- AMR codec [1] commonly used for narrowband speech in mobile networks. The audio bandwidth covers frequencies up to 4 kHz. Four bit rate modes were evaluated: 4.75 kbit/s, 7.95 kbit/s, 10.2 kbit/s, and 12.2 kbit/s.
- AMR + ABE: AMR codec followed by ABE processing. Four combinations were evaluated: AMR at 7.95 kbit/s and 12.2 kbit/s followed by ABE1 and ABE2.
- AMR-WB codec [3] for wideband speech, currently being deployed in an increasing number of mobile networks [2]. The audio bandwidth extends up to 7 kHz. Four bit rate modes were evaluated: 6.6 kbit/s, 8.8 kbit/s, 12.65 kbit/s, and 23.85 kbit/s.
- AMR-WB + SWB-ABE: AMR-WB codec followed by SWB-ABE processing. Two bit rate modes of AMR-WB were evaluated: 12.65 kbit/s and 23.85 kbit/s.
- Opus [4], a real-time, variable and fixed bit rate codec with the highest voice quality currently available in open source. Four constant bit rates (CBR) were evaluated. The corresponding bandwidths were selected by the codec based on bit rate: 10.2 kbit/s (narrowband, 4 kHz), 12.65 kbit/s (mediumband, 6 kHz), 16 kbit/s (wideband, 8 kHz), and 20 kbit/s (superwideband, 12 kHz).
- ITU-T G.722.1 Annex C [5], a low-complexity superwideband voice codec widely deployed in video teleconferencing services. The audio bandwidth is 14 kHz. Two bit rate modes were evaluated: 24 kbit/s and 32 kbit/s.
Two bit rate modes with 14-kHz audio bandwidth were evaluated: 28 kbit/s and 40 kbit/s.

3.1. Listening tests

Three tests were arranged with different background noise conditions and highpass filter cutoff frequencies. All speech samples were filtered with a highpass filter having a flat response in the passband and a cutoff frequency of 150 Hz (test 1) or 50 Hz (tests 2 and 3). The 150-Hz cutoff corresponds to the response of a mobile phone in the far end where low-frequency noise is reduced by highpass filtering. In practice, low frequencies are attenuated also if a mobile phone is used in the near end because the low-frequency reproduction capability of an earpiece is typically very limited. On the other hand, codec characterization tests commonly employ a highpass filter with a cutoff of 50 Hz and thus minimal limitation of the passband at low frequencies. Since ABE quality is known to vary from talker to talker, short speech samples were chosen in tests 1 and 2 so that 8 talkers could be included.

- **Test 1**: Clean speech, highpass cutoff 150 Hz, 8 talkers (4 females, 4 males), sentence pairs of about 6 seconds.
- **Test 2**: Clean speech, highpass cutoff 50 Hz, 8 talkers (4 females, 4 males), single sentences of about 4 seconds.
- **Test 3**: Noisy speech, highpass cutoff 50 Hz, 4 talkers (2 females, 2 males), sentence pairs of about 7 seconds. Four noise types: car noise with signal-to-noise ratio (SNR) of 15 dB, street noise (SNR 15 dB), cafeteria noise (SNR 20 dB), and office noise (SNR 20 dB).

Modified ACR tests were used for evaluation. Instead of the 5-point scale defined in [14], a discrete 9-point scale was used and only the extreme categories (1 “very bad” and 9 “excellent”) were labeled with verbal descriptions [27].

The tests were arranged in the listening test laboratory of Nokia Research Center [35]. Subjects were seated in soundproof booths and listened to samples diotically (the same signal to both ears) through an RME Multiface II audio interface and Sennheiser HD-650 headphones. The listening level was set to a sound pressure level (SPL) of 76 dB and could not be changed by the listeners. Listeners heard each test sample once (no relisting allowed) and gave their opinion using a discrete 9-step scale. A training session with 12 samples preceded each test. Twenty-eight listeners participated in each test. In all the tests, 5–6 of the participants were expert listeners (35–47 years of age) working in the field of audio signal processing. The remaining participants were naive listeners (15–23 years of age).

4. Results

The mean opinion scores on the 9-point scale (MOS9) and 95-percent confidence intervals of all three tests are shown in Figure 2. Additionally, the mean scores and 95-percent confidence intervals of AMR, AMR-WB, ABE, and Opus conditions are presented in Figure 3 as a function of codec bit rate.

Two-tailed independent-samples t tests were conducted to compare the mean scores within each test. Statistically significant differences (α = 0.05) between ABE conditions and the conditions used as input to ABE are presented in Table 1. For clean speech and 150-Hz highpass filtering, all ABE conditions were significantly better than the corresponding reference conditions. For clean speech with 50-Hz highpass filtering, all NB-
to-WB ABE conditions were significantly better than the reference conditions and the improvement by SWB-ABE following AMR-WB at 23.85 kbit/s was close to statistical significance ($p = 0.064$). There were no significant differences between ABE conditions and the corresponding reference conditions in the test with noisy speech. Also, no significant differences were found between ABE1 and ABE2 in any of the tests.

Table 1: Statistically significant differences between ABE conditions and the corresponding reference conditions. In each case, condition 2 is the same codec as condition 1 followed by the indicated ABE method. $df = 446$ in all these cases.

<table>
<thead>
<tr>
<th>condition 1</th>
<th>MOS9</th>
<th>condition 2</th>
<th>MOS9</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean speech, 150-Hz highpass</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMR 7.95</td>
<td>3.46</td>
<td>ABE1 3.88</td>
<td>2.61</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td>AMR 7.95</td>
<td>3.46</td>
<td>ABE2 3.89</td>
<td>2.63</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td>AMR 12.2</td>
<td>3.75</td>
<td>ABE1 4.20</td>
<td>2.74</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>AMR 12.2</td>
<td>3.75</td>
<td>ABE2 4.27</td>
<td>3.18</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>AMR-WB 12.65</td>
<td>5.78</td>
<td>SWB-ABE 6.11</td>
<td>2.24</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td>AMR-WB 23.85</td>
<td>5.70</td>
<td>SWB-ABE 6.12</td>
<td>2.74</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>Clean speech, 50-Hz highpass</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMR 7.95</td>
<td>3.48</td>
<td>ABE1 3.89</td>
<td>2.64</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td>AMR 7.95</td>
<td>3.48</td>
<td>ABE2 3.98</td>
<td>3.14</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>AMR 12.2</td>
<td>3.87</td>
<td>ABE1 4.49</td>
<td>3.65</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>AMR 12.2</td>
<td>3.87</td>
<td>ABE2 4.44</td>
<td>3.42</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>

5. Conclusions

Two NB-to-WB ABE methods (ABE1 and ABE2) and one WB-to-SWB ABE method (SWB-ABE) were evaluated in subjective listening tests together with standardized speech codecs with different audio bandwidths. The ABE methods were designed to be implementable in real time with reasonable delay and resources. Evaluations were organized as ACR listening tests commonly used for the quality characterization of speech codecs. Tests were arranged for both clean and noisy speech, and two clean-speech tests were organized with different highpass cutoff frequencies: 50 Hz and 150 Hz. In each test, ABE methods were applied to speech coded with the AMR and AMR-WB codecs using two different bit rates.

For clean speech, NB-to-WB ABE methods were found to significantly improve the speech quality. For noisy speech, no statistically significant improvement was obtained, but the mean scores of NB-to-WB ABE methods were slightly higher than those of the corresponding narrowband cases. Differences in scores between ABE1 and ABE2 were negligible except for the noisy case, where ABE2 was scored slightly but not significantly better. The benefit of the WB-to-SWB ABE was smaller. A statistically significant improvement for SWB-ABE was reached only for clean speech with 150-Hz highpass filtering. For noisy speech, the mean scores of SWB-ABE were close to those of the wideband reference conditions. Overall, the results for 50-Hz and 150-Hz highpass filtering were similar except that the scores were generally slightly higher in the test with a 50-Hz cutoff.

The results are in line with many earlier studies on ABE showing that NB-to-WB ABE improves speech quality [16, 19, 9]. On the other hand, the results contrast with those presented in [23] where none of the ABE methods significantly improved the speech quality. A possible reason for this difference is the use of the IRS send filter in [23] instead of a flat magnitude response in the passband, which corresponds more closely to the characteristics of today’s mobile devices and digital networks.

NB-to-WB ABE methods improved the mean scores in all cases including noisy speech and different codec bit rates. WB-to-SWB ABE also improved the mean scores for clean speech and had no practical effect on the mean scores for noisy speech. The results support the feasibility of ABE in varying use cases including different codec bit rates, highpass filtering cutoffs, and downlink noise conditions. ABE has also been shown to improve the intelligibility, which was not evaluated in this study.
6. References


