Evaluating artificial bandwidth extension by conversational tests in car using mobile devices with integrated hands-free functionality

Laura Laaksonen, Ville Myllylä, Riitta Niemistö

Nokia, Symbian Smartphones, Audio Technology, Finland

laura.laaksonen@nokia.com, ville.myllyla@nokia.com, riitta.niemisto@nokia.com

Abstract

This paper describes an artificial bandwidth extension (ABE) method that generates new high frequency components to a narrowband signal by folding specifically gained subbands to frequencies from 4 kHz to 7 kHz, and improves the quality and intelligibility of narrowband speech in mobile devices. The proposed algorithm was evaluated by subjective listening tests. In addition, rarely used conversation test was constructed. Speech quality of 1) narrowband phone call, 2) wideband phone call, and 3) narrowband phone call enhanced with ABE were evaluated in conversational context using mobile devices with integrated hands-free (IHF) functionality. The results indicate that in IHF use case, ABE quality overcomes narrowband speech quality both in car noise and in quiet environment.

Index Terms: speech enhancement, artificial bandwidth extension, conversation test

1. Introduction

Most current cellular networks transmit speech signals in narrowband format. The limited frequency band from 300 Hz to 3.4 kHz reduces both speech quality and intelligibility. Adaptive multirate wideband (AMR-WB) codec [1] with wider speech bandwidth of 50–7000 Hz was introduced already years ago but the deployment of wideband speech has been slow. Although the number of mobile terminals with wideband support in the market is gradually increasing, only few operators around the world offer wideband speech transmission in their networks. Therefore, narrowband and wideband calls will co-exist for years and during that time the end users will suffer from a quality gap between narrowband and wideband speech. Especially, when handovers between wideband and narrowband networks during a call become common, mobile device users will suffer from annoying voice quality alteration. This is because after a handover from wideband to narrowband, the quality of narrowband speech is perceived worse than usually for some time [2]. The slow emergence of wideband speech coding has increased the interest towards artificial bandwidth extension methods that can be used to enhance narrowband speech quality. ABE methods aim to regenerate the missing frequency content in the high band, i.e. frequency range from 4 to 8 kHz. The extension is typically performed without any transmitted side-information in, for example, the receiving mobile terminal.

Many artificial bandwidth extension methods have been introduced during the last decades. The performance of these methods has been mostly evaluated by objective measures or by subjective listening tests where participants are placed in a listening situation, i.e. tests having only listening context. In such tests, pre-recorded speech samples are usually used, and the subjects are able to listen to the test samples as many times as they want. The content of the signal becomes easily irrelevant, whereas, in conversational situations of everyday life, the mobile device user is paying attention to both the form of the signal, i.e. the acoustic signal, and the content of the signal, i.e. the semantic information [3]. In other words, in subjective listening tests listeners concentrate to assess quality – not intelligibility of speech. Moreover, conversation includes both listening and talking periods, not only listening periods. These periods alternate according to the interaction with the interlocutor. The International Telecommunication Union describes in the Recommendation P.805 conversation tests that can be used to evaluate communication quality in a more realistic situation simulating the actual service conditions experienced by telephone customers [4]. In such tests two participants have a real-time conversation through a transmission chain, and they give their opinion on the quality. Conversational tests are, in general, time consuming and, thus, expensive to arrange. Therefore, they are rare in the literature, and also rare in the field of ABE research.

The ABE method introduced in this article generates new high frequency components to a narrowband signal by folding specifically gained subbands to frequencies above 4 kHz. The algorithm development was a continuation for the ABE method discussed in [5]. The objective was to improve the extension of fricatives and to implement the signal path completely in the time domain. The performance of the proposed algorithm was first examined by listening tests. Then, conversational tests were conducted to evaluate the effect of ABE processing in telephone communication using mobile devices with IHF functionality in car environment. The aim was to assess whether the occasional artifacts produced by ABE algorithm are disturbing in conversational context or not. Moreover, it was anticipated that the ABE processing would be especially beneficial to IHF use case. The ABE processing increases loudness by adding new energy to the higher frequencies. That should help with tiny IHF loudspeaker(s), which cannot typically provide enough volume for noisy conditions such as a car interior, where the hands-free phone usage is often mandatory by law. Also, since the narrowband information is kind of duplicated above 4 kHz, the generated higher frequency content will be less masked by typical low frequency concentrated noise.

2. Algorithm

A flow diagram of the method is shown in Figure 1. The input to the ABE algorithm, s_{nb}, is a 10 ms frame of narrowband signal with sampling frequency of 8 kHz. The signal is interpolated to 16 kHz sampling frequency and lowpass filtered. This output contains the original narrowband signal in the frequencies of 0–4 kHz.

The content of the narrowband frame is analyzed by calculating several time domain and frequency domain features. In addition, background noise levels of the input signal, s_{nb}, and the microphone signal of the mobile device, s_{mic}, are estimated. Based on these features and noise estimates, each frame is classified either as a voiced frame, a plosive or a sibilant.

Copyright © 2011 ISCA

In order to create new content to higher frequency range of 4–8 kHz, the narrowband signal is divided into four subbands by a nondecimating analysis filterbank. Specific gains based on the classification and noise levels, are computed for each subband. The gains are applied to the subbands and the subbands are combined back together by summing them up. Finally, the new high band is created by folding the specifically gained and summed subbands to the high band by applying interpolation and highpass filtering. The resulting new high band, with 16 kHz sampling frequency, is added to the interpolated and lowpass filtered narrowband signal. As a result, the artificial wideband signal, s_{abe}, is obtained.

In the next subchapters, the new blocks of the algorithm are discussed in more detail.

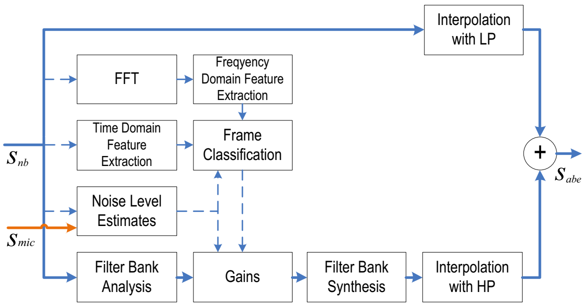

Figure 1: *Flow diagram of the implemented algorithm.*

2.1. Filterbank

Nondecimating cosine-modulated filterbank is designed using Mth-band filter ($M = 10$) as a prototype filter. The filter minimizes a constrained least squares criterion

$$\int \left| e^{-j\tau\omega} - H(\omega) \right|^2, \; h_i = h_{-i}, i > 0, h_{nM} = \begin{cases} \frac{1}{M}, \mathrm{n} = 0 \\ 0, \mathrm{n} \neq 0 \end{cases} \quad (1)$$

over frequency range ω with respect to coefficients h_i of $H(\omega)$. The group delay $\tau = 17$ is the delay produced by the filterbank and dictates the order of the filters. The filterbank is used as analysis filterbank, and on synthesis side the bands are summed up. The bands corresponding to highest frequencies (above 7 kHz) get weight zero; therefore, those bands or the corresponding filters are not present in the practical implementation. Figure 2 illustrates the filterbank.

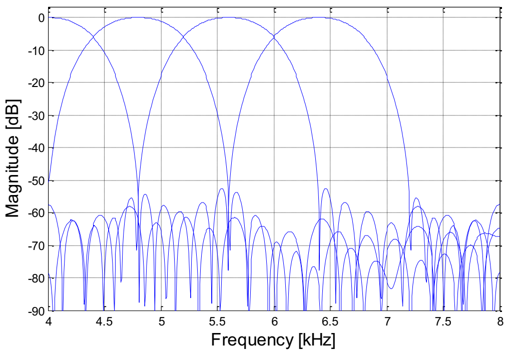

Figure 2: *Filterbank used in weighting the high band.*

2.2. Narrowband features and frame classification

In the proposed method the frames are classified into three categories (voiced frames, plosives and sibilants). The classification is based on k-means algorithm where each cluster is represented by a cluster centroid that is a feature vector. The feature vector is computed from the narrowband signal. For each frame, the distances from the centroids are computed and a cluster with the smallest distance is selected.

The feature set for frame clustering consists of the following features:

- Gradient index, defined as

$$x_{gi} = \frac{\sum_{\kappa=1}^{N_{\kappa}-1} \Psi(\kappa) |s_{nb}(\kappa) - s_{nb}(\kappa-1)|}{\sqrt{\sum_{\kappa=0}^{N_{\kappa}-1} \left(s_{nb}(\kappa)\right)^2}} \quad (2)$$

  where κ is the sample index, $\Psi(\kappa) = 1/2\,|\psi(\kappa) - \psi(\kappa-1)|$ and $\psi(\kappa)$ is the sign of the gradient $s_{nb}(\kappa) - s_{nb}(\kappa-1)$.

- Energy ratio, as a ratio between the energy of the current frame and the energy of the previous frame.
- Spectral centroid of gravity is computed from the 128-tap FFT spectrum and defined as:

$$x_{scg} = \frac{\sum_{i=0}^{N_i/2} i \left| s_{nb}\left(e^{j\Omega_i}\right) \right|}{\left(\frac{N_i}{2}+1\right) \sum_{i=0}^{N_i/2} \left| s_{nb}\left(e^{j\Omega_i}\right) \right|} \quad (3)$$

  where $s_{nb}\left(e^{j\Omega_i}\right)$ is the i-th Fourier coefficient and $N_i = 128$ is the length of the Fourier transform.

The features were normalized to unit variance before training the k-means centroids with Finnish speech from three female and three male speakers. The training data consisted of 1930 10 ms frames of speech that were labeled by hand.

2.3. Gains

After the classification of a frame into either a voiced frame, plosive frame, or sibilant frame, the gains for weighting the subbands on high band are computed. The gain calculation is based on frequency domain control points that were used in shaping the high band in [2]. The frequency domain control points define the amount of gain needed to shape the folded narrowband spectrum in the high band at frequencies 4 kHz, 5 kHz, 6 kHz, 7 kHz and 8 kHz. The control points for frame *n* are computed from predefined formulas of the form:

$$C_{k,f} = b_{k.f} + a_{k,f} \cdot x_{ns}(n), \quad (4)$$

where $b_{k,f}$ and $a_{k,f}$ are control point constants for frame cluster *k* at frequencies *f*. Parameter x_{ns} is a narrowband slope that quantifies the slope of the narrowband spectrum in the frequency range 300–3000 Hz. It is computed from the the narrowband FFT spectrum. The predefined formulas were optimized offline using a genetic algorithm (GA) –based search [5].

In the proposed ABE algorithm the gains for the subbands are computed from the control points using the following mapping:

$$\begin{aligned} g_{band1} &= C_{\mathrm{k},4kHz} \\ g_{band2} &= C_{\mathrm{k},5kHz} \\ g_{band3} &= \tfrac{1}{2}\left(C_{\mathrm{k},5kHz} + C_{\mathrm{k},6kHz}\right) \\ g_{band4} &= \tfrac{1}{2}\left(C_{\mathrm{k},6kHz} + C_{\mathrm{k},7kHz}\right) \end{aligned} \quad (5)$$

Some further noise dependent control was added to these gains. For noisy narrowband signals, the whole high band is attenuated to prevent the extended noise in the signal from becoming annoying. On the other hand, if the ambient noise around the listener, i.e., the noise level in the microphone signal, increases, the high band can be carefully further amplified, because noise masks possible artifacts. This kind of ambient noise dependent control increases intelligibility of the ABE output.

2.4. Listening tests

Subjective listening tests were organized to evaluate the performance of the proposed algorithm. Comparison category rating (CCR) test [6] was chosen as an evaluation method. The test included five different processing types that were all compared with each other. The test material included high quality Finnish speech recordings with 16 kHz sampling frequency from two female and two male speakers. Two speech samples were selected from each speaker resulting in total eight different speech samples, each containing one spoken sentence with duration of approximately two seconds.

The signal level of the test samples was adjusted to -19 dBov. Office noise with SNR 35 dB was added to test signals. The test signals were filtered with GSM1 filtering [7] to mimic frequency response of a typical GSM transmission path. The narrowband reference was down-sampled and processed twice through AMR NB encoding and decoding (with bit rate of 12.2 kbit/s). The wideband reference was processed through AMR-WB encoding and decoding (with bit rate of 12.65 kbit/s). In addition to the narrowband and wideband references, also a bandlimited wideband reference was used. The bandlimited wideband signal had roughly the same bandwidth as the artificial bandwidth extended signals and it was obtained by bandstop and lowpass filtering the wideband reference. Two artificial bandwidth extension versions were included in the test. The narrowband reference was used as an input for the algorithms. The method proposed in this paper is here referred to as ABE, and a reference algorithm from [5] as ABE-ref. Average magnitude spectra of one of the sentences from the listening test for each processing type are shown in Figure 3.

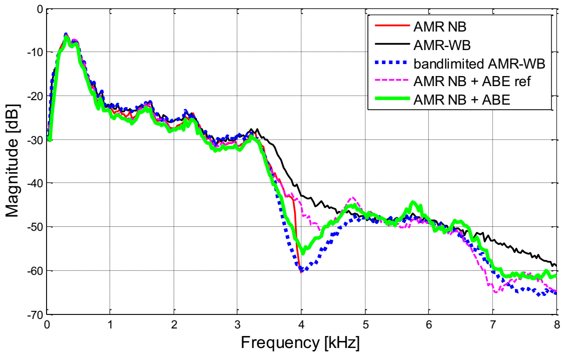

Figure 3: *Example of average magnitude spectra of speech for each processing type.*

The listening test was arranged in a listening room. Ten listeners with age from 31 to 47 participated the test. The listeners listened to the samples with Sennheiser HD 600 headphones and they were able to listen to each sample pair as many times as they wanted. Each listener listened to 80 pairs of test samples. In addition, 10 null pairs were included in the test. The listeners were asked to grade the overall quality of the latter sample compared to the first sample, using a seven-point scale from -3 to 3 called comparison mean opinion score (CMOS) [6].

The results of the test are shown in Figure 4. The best score (1.43) was given for the AMR-WB reference, whereas the lowest score (-1.71) was given for the AMR NB reference. The bandlimited version of the AMR-WB was rated the second best with score 0.89. The scores for the proposed ABE and the reference ABE were -0.34 and -0.27, respectively.

According to the test results, the benefits of ABE processing are obvious, even though the ABE quality is closer to narrowband than to wideband. The proposed ABE method and the reference algorithm obtained nearly the same score from the listening test. This can be explained by the fact that the gain calculation is based on the control points of the reference algorithm. However, the slightly better score of ABE method suggests that the k-means based frame clustering improves the classification of sibilants. In fact, based on expert listening and comparison between ABE processed samples and ABE ref processed samples, 30% of the sibilants sounded brighter and had more energy in the highband. In addition, the filterbank based processing resulted in less distortion than the FFT based implementation in [5].

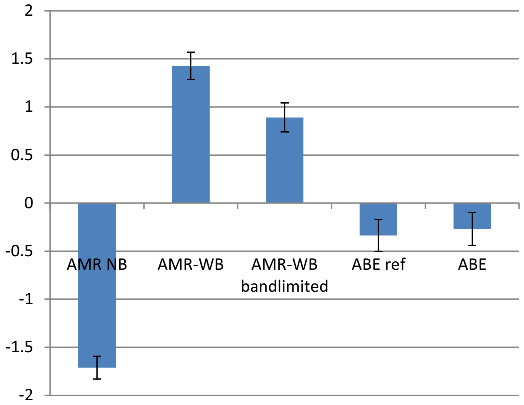

Figure 4: *Relative scores of the processings in the CMOS listening test with 95 % confidence intervals.*

2.5. Car tests

Conversational test setup was constructed to evaluate artificial bandwidth extension using true telecommunications system. In the test, paired sets of test subjects held conversations through a test 3G network using mobile devices with integrated hands-free (IHF) functionality. The test setup was asymmetric; the test participant was sitting in a car in garage, whereas the other participant, who had an assisting role in the test, was in an office room.

Three mobile phones of same model were placed in front of a wheel in a test car (Ford Mondeo), as shown in Figure 5. The frequency responses, of the IHF speakers, measured with the phone on a desk are illustrated in Figure 6. The listener sat on the driver's seat and had a conversation with another person sitting in an office room. There were two open phone call lines during each task. The speaker in the office room talked to two mobile devices at the same time and the listener in the car was supposed to freely switch between two devices while discussing with the other speaker.

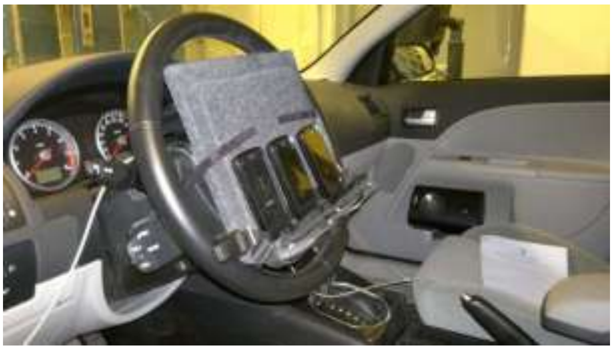

Figure 5: *The mobile phones were placed in front of the wheel in the test car.*

Three different phone call modes were included in the test: 1) conventional narrowband (NB) phone call, 2) wideband (WB) call with AMR-WB, and 3) ABE call with narrowband coding and ABE processing applied in the device. The three mobile devices were tested in pairs with all combinations. During first three calls, car noise was simulated in the car cabin using SoundCar - Multidimensional sound playback in a real vehicle [8], whereas the last three calls were made in quiet.

Ten expert listeners participated in the test. Everyone of them had six short telephone conversations with the conversationalist in the office room. The duration of the conversations was approximately 2–5 minutes. Four speakers (two female and two male) assisted the test by acting as a conversationalist in the office room. All the conversations were conducted in Finnish.

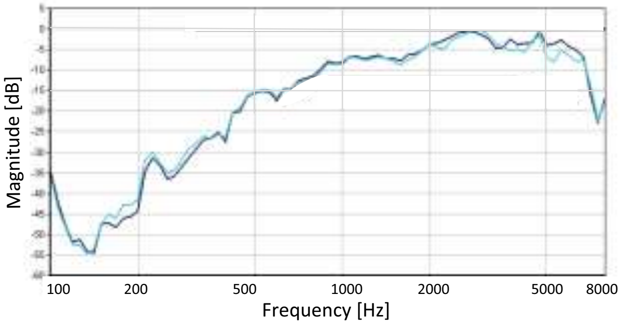

Figure 6: *Frequency responses of two mobile devices used in the car test.*

The listeners were asked to compare the voice quality of two mobile devices with IHF functionality activated, and to decide which one he/she preferred. In addition, they were asked to briefly reason their opinion. The test subjects were not told of the nature of the effect under test nor which device had which processing. In practice, the IHF functionality were activated in both devices having a call, but, in turns, one of them was put on hold, i.e. microphone was muted and speaker silenced. The conversation between the test person and the talker in the office room was based on conversation scenarios that included, for example, hotel room and theatre inquiries. The conversation scenarios were biased so that the talker in the office room talked more than the listener in the car by giving long answers to the inquiries, but still the tasks were conversational.

The results from the test are shown in Figure 7. In the comparison between NB and ABE, 80% of the listeners preferred ABE over NB. WB was better than NB with preference score of 90%, and WB was always preferred over ABE.

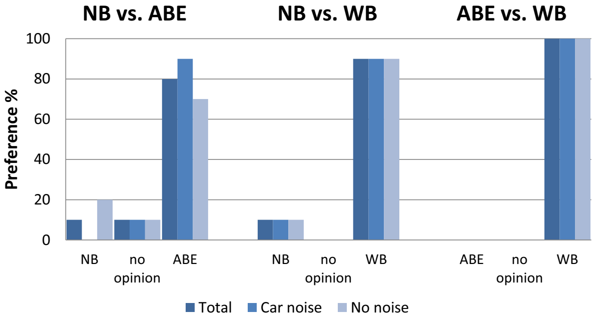

Figure 7: *Preference percentages of the comparisons in the car test.*

The loudness of the mobile devices was adjusted subjectively in car noise to a level where the listener had to focus on listening in order to understand the talking partner. The loudness level was kept the same for all devices, listeners and all conversation tasks. In car noise, when reasoning why ABE was preferred over NB, seven of the ten listeners commented that the ABE enabled phone was more intelligible and easier to listen.

3. Conclusions

Artificial bandwidth extension (ABE) algorithms can be seen as quality and intelligibility enhancement methods for narrowband speech. Along with wideband speech transmission, the role of ABE methods is more and more to decrease the quality and intelligibility gap between narrowband and wideband speech.

The introduced ABE method generates new high frequency components to a narrowband signal by folding specifically gained subbands to frequencies from 4 kHz to 7 kHz. The method was evaluated by CCR listening test indicating clear quality improvement compared to narrowband speech.

Subjective evaluation between 1) narrowband, 2) wideband and 3) narrowband with ABE processing phone calls were carried out in conversational context in car environment. The conversation test was conducted with true mobile devices using integrated hands-free (IHF) functionality. The positive results from the car test support the preconception that artificial bandwidth extension is especially useful in IHF use case. Intelligibility is improved due to both wider bandwidth and increased loudness compared to narrowband speech. Furthermore, the conversational context in the subjective testing seems to give beneficial and novel information about ABE methods. When the user is concentrating on the conversation, small occasional artifacts are not perceived, or at least are not disturbing. For future work, more comprehensive tests could be arranged for making more detailed analysis of the effect of the conversational context on ABE methods.

4. References

[1] 3rd Generation Partnership Project (3GPP), "AMR wideband speech codec; general description, 3GPP TS 26.171," 2001, version 5.0.0.

[2] 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. On Consumer Electronics, vol. 55, no. 2, May 2009.

[3] M. Guéguin, R. Le Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, and V. Barriac, "On the evaluation of the conversational speech quality in telecommunications," EURASIP Journal on Advances in Signal Processing, 2008.

[4] ITU-T recommendation P.850, methods for objective and subjective assessment of quality, Int. Telecommun. Union, 2007.

[5] H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen, and P. Alku, "Evaluation of an artificial bandwidth extension method in three languages," IEEE Trans. Audio, Speech, and Language Processing, vol. 16, no. 6, pp. 1124-1137, Aug. 2008.

[6] ITU-T recommendation P.800, methods for subjective determination of transmission quality, Int. Telecommun. Union, 1996.

[7] ITU-T recommendation G.191, software tools for transmission systems, Int. Telecommun. Union, 2010.

[8] SoundCar - Multidimensional sound playback in a real vehicle, http://www.head-acoustics.de/eng/nvh_soundcar.htm.