BINAURAL ARTIFICIAL BANDWIDTH EXTENSION (B-ABE) FOR SPEECH

Laura Laaksonen, Jussi Virolainen

Nokia Devices, Audio R&D
P.O.Box 407, FI-00045 Nokia Group, Finland
laura.laaksonen@nokia.com, jussi.virolainen@nokia.com

ABSTRACT

Efficient teleconference system can be created by extending the speech bandwidth and using spatial audio. If wideband speech transmission was not supported as such, an alternative way to extend the bandwidth is by exploiting an artificial bandwidth extension (ABE) algorithm. In such methods the audio bandwidth is extended in the receiving end, without any transmitted extra information. When considering spatial audio it should be noted that ABE methods have been developed for monaural signals and they cannot be applied as such to binaural signals due to a possible mismatch of binaural cues in the created frequency range. In this paper binaural ABE, B-ABE, is proposed. Subjective listening tests were used to evaluate the proposed method. The results showed that the localization information is preserved well in B-ABE processing.

Index Terms— Artificial bandwidth extension, HRTF filtering, spatial audio, teleconferencing.

1. INTRODUCTION

Speech communication in teleconferencing can be made more effective by extending the speech bandwidth and using spatial audio. When the remote conference participants are placed at different virtual positions around the listener, the ”cocktail party effect” can be exploited [1]. It means listener's ability to focus one's listening attention on a certain speaker when others are talking simultaneously or when background noise is present. Spatial representation of speech sources has many advantages, it improves speech intelligibility and makes speaker detection and separation easier. Furthermore, reduced listening effort and more naturally sounding conference environment may prevent listening fatigue.

Stereo audio is already applied in commercial high quality group video conferencing systems. Spatial audio can be beneficial in mobile audio conferencing solutions as well. A promising approach is based on a centralized conferencing architecture, in which spatial processing and mixing take place in the conference bridge. Each terminal sends a monophonic signal to the bridge, which sends a binaural signal back to the terminal to be reproduced through stereo headphones.

The spatialization (or 3D processing) of a monophonic input signal can be done by applying Head Related Transfer Function (HRTF) processing to produce binaural signal that is suitable for headphone reproduction. To simplify, the mono sound source is panned by modifying both delay and amplitude of the left and right channel. The delay between the two channels, called Interaural time difference (ITD), models the time delay in the arrival of the sound signal between the right and left ear. Whereas the level difference, called interaural level difference (ILD) models the frequency dependent intensity difference between the two ears. ITD is the dominant localization cue below 1.5 kHz and ILD in the higher frequency range.

Today, speech is still mostly transmitted in narrowband. For example, in PSTN and GSM networks, the supported bandwidth is from 0.3 kHz to 3.4 kHz. In the future, wideband speech transmission will become more prevalent but the transition phase requires a significant amount of effort from the operators and end users, since the whole transmission chain including terminals and network elements should support wideband transmission. It is known [2] that the bandwidth plays an important role in the performance of 3D audio. In general, better intelligibility and better localization is achieved with wider bandwidth. During the transition phase from narrowband to wideband speech, the quality and intelligibility of narrowband speech can be enhanced by extending the bandwidth of the received signal artificially. That is, the information obtained in the narrowband signal is used to create new content to higher frequency range in the receiving end of the transmission. These method are usually referred to as artificial bandwidth expansion (ABE) or bandwidth extension (BWE) methods.

So far, all the ABE methods have been developed for monaural speech signals. We refer to these methods as monophonic ABE. Most of these methods are based on a source filter model, in which the speech production process is modelled by an excitation signal and a vocal tract filter that modifies the spectrum of the excitation signal [3]. In such methods, the excitation signal and the highband (4-8 kHz) envelope are estimated separately [4] [5]. The proposed Binaural-ABE (B-ABE) is based on a monophonic ABE method that uses spectral folding and adaptive shaping curves in the frequency

domain to create the missing data in the highband [6].

If a monophonic ABE is applied for extending the left and right channels of a binaural signal separately, binaural artefacts are possibly created. Due to the temporal and spectral differences between the two narrowband channels, there is no guarantee that the created highbands would match in terms of binaural cues. Especially, if frame based processing is applied, a separate processing of the left and right channels may cause significant differences to highband that are not aligned with the ILD and ITD of the narrowband. For example, fricatives have a considerable amount of energy above 4 kHz, and if the binaural cues in the highband are inaccurate, the listener would hear the fricative from another direction than the rest of the word. As a result, there is a need for binaural-ABE (B-ABE) method that extends the bandwidth of binaural speech signals and creates adequate binaural cues to the highband. The challenge in B-ABE processing is to preserve the localization information of the narrowband binaural signal. The positions of the talkers should not change or fluctuate despite the bandwidth change. In addition, possible artefacts of the artificial bandwidth extension should not degrade the localization information of the signal.

2. METHOD

Our conference bridge and terminal with B-ABE function is presented in Fig. 1. Mono input signals from the terminals are spatialized and mixed in the conference bridge. The binaural signal is transmitted to a terminal where B-ABE processing is performed. First, the time and level difference between the left and right channels are estimated from the narrowband binaural signal. The bandwidth extension then utilizes the ILD and ITD estimates to match the binaural cues in the created highband of the binaural output signal. The purpose of the processing is to spatialize artificially generated highband to the same location as the original sound source.

There can be either one or multiple simultaneous speakers in a spatialized signal at a time. During only one talker, the spatialization of the highband is straightforward, since the level and time differences between the channels are somewhat unambigous. Whereas, during simultaneous speech, the speakers might be virtually placed to different positions around the listener. In such situation, the ITD and ILD vary between the speakers and the spatialization of the highband has to be made based on one dominant speaker at a time. For example, the speaker with highest intensity can be selected as a dominant speaker.

3. IMPLEMENTATION OF THE METHOD

The detailed implementation of the binaural artificial expansion method shown in Fig. 2 is based on a monophonic ABE algorithm presented in [6].

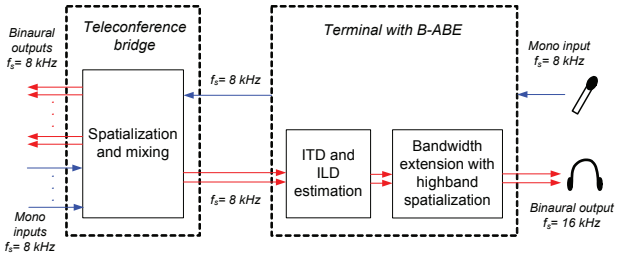

Fig. 1. Description of a system including a conference bridge and a terminal with B-ABE function. 3D processing is performed in the conference bridge and the binaural signal is sent to the terminal where B-ABE processing takes place.

The left and right channels of the binaural input signal are processed in cascade in 10 ms frames and an overlapping trapezoidal windowing is used to prevent the aliasing effect. The ITD between the right and left channels is estimated using average magnitude difference function (AMDF), defined as:

$$d(i) = \frac{1}{N}\sum_{k=1}^{N} |x_l(k) - x_r(k-i)| \quad (1)$$

where x_l is the left channel, x_r is the right channel, N is the analysis frame length, and i is the delay in samples. The ITD is the value of delay, i, which results in the minimum value of the following set of magnitude differences:

$$min\{d(-10), d(-9), \ldots d(10)\} \quad (2)$$

where 10 samples corresponds to ITD of 1.25 ms at sampling frequency of 8 kHz. The ILD of true wideband binaural signals is frequency dependent but in this implementation a constant level difference is applied to the highband. A simple linear approximation is used to code the ITD estimate into level difference. The level difference in dB is $l = \alpha i$, where α is a parameter that is estimated from the HRTFs.

The left channel is processed with a monophonic ABE. Since the method modifies the highband in the frequency domain, FFT and IFFT processing are needed. Overlap-add technique is used to concatenate the frames into a continuous signal again.

A copy of a frequency domain presentation of the created highband is extracted from the left signal path for further processing. First, the narrowband part of the FFT buffer is set to zero, and the level difference of left and right channels is set to the target level l. Wideband IFFT is computed and the estimated ITD is created in the time domain. After overlap-add procedure, the signal contains the new highband for right channel.

The narrowband part of the right channel has to be interpolated to 16 kHz sampling frequency, before summing the

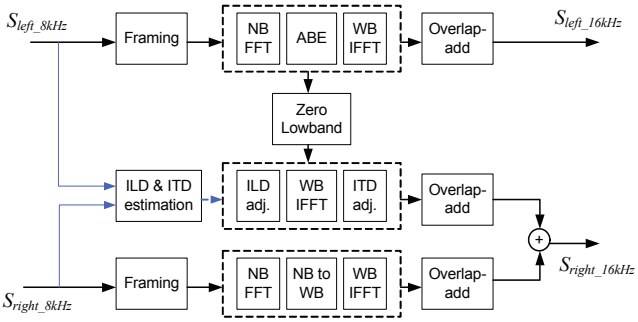

Fig. 2. Block diagram of the implementation of the B-ABE algorithm. The input signal consists of narrowband channels s_{left_8kHz} and s_{right_8kHz}. The artificially extended binaural signal consists of channels s_{left_16kHz} and s_{right_16kHz}.

narrowband and highband parts together. The interpolation is implemented in the frequency domain in order to avoid unmatched group delays in cascade signal paths. Finally, the output of the right channel is obtained by summing the narrowband and highband signals together.

4. LISTENING TEST

4.1. Test Arrangements

The performance of the binaural artificial expansion method was evaluated with subjective listening tests. The tests were designed to measure how is the localization information in a narrowband binaural signal preserved in the B-ABE processing. The listening test methods used in this study were two separate degradation category rating (DCR) tests. The task of the listener was to evaluate the localization information of the second sample compared to the first sample using the following degradation mean opinion score (DMOS) scale: difference/degradation is: 5, inaudible; 4, audible but not annoying; 3, slightly annoying; 2, annoying; 1, very annoying.

Test samples in the first test, test A, included speech of one speaker, whereas in the second test, test B, simultaneous speech of two speakers were used. Total 36 different Finnish sentences were chosen from the NTT database [7]. Three different binaural processings were included in the tests, 3D narrowband (3D NB), 3D wideband (3D WB) and B-ABE. The original samples with 16 kHz sampling frequency were first filtered with a model of the input characteristics of a mobile station, which is basically a high-pass filter with the cutoff frequency at about 200 Hz. The 3D wideband signal was generated directly by 3D processing, which included direct sound filtering using HRTF set that has been calculated for a hard head with a moderately absorbing torso [8]. The 3D narrowband signal was produced by downsampling the signal to 8 kHz and applying then 3D processing. The 3D narrowband signal served as an input to B-ABE processing.

Test A included two rehearse sample pairs and 36 actual test sample pairs. Total 12 different speech samples (2 from 3 female and 3 male speakers) were processed. B-ABE processings were evaluated in reference to 3D WB and 3D NB processings. In addition, 3D NB was evaluated in reference to 3D WB samples. In four of the samples the speaker was in front of the listener, i.e. in the direction of $0°$, in four of the samples the speaker was in the direction of left/right $45°$ and in four of the samples the speaker was in the direction of left/right $90°$. The listeners were asked to compare the localization of the second sample to the localization of the first samples. The question to the listeners was: Can you hear any difference in the localization of the speaker?

Test B also included two rehearse sample pairs and 36 actual test sample pairs. All the samples consisted of two simultaneous speakers. The samples were spoken by the same 6 speakers as in test A. Each test sample contained two sentences, the second sentence started 0.6 s after the first one. The simultaneous sentences were spoken by two different speakers of either the same or different gender. The virtual locations of the two simultaneous speakers were always different and all the possible combinations of directions of $0°$, left/right $45°$ and left/right $90°$ were used. The listeners were asked to compare the stability of the localization in the second sample compared to the first one. The questions to the listeners were: Is there any fluctuation in the localization of the speakers especially during simultaneous speech? If there are any distortion/artefacts, are they coming from different directions as the actual speech?

Total 10 expert listeners (4 female, 6 male, ages between 29 and 38) participated the test. No information on the purpose of the test were given to the listeners. All the listeners were native Finnish speakers. The samples were listened through high quality headphones.

4.2. Results

The listeners' opinions on the localization of the B-ABE processed samples compared to 3D narrowband and 3D wideband references, and 3D narrowband samples compared to 3D wideband reference, were investigated. The results from the test A are shown in the left panel of Fig. 3. In listeners' opinion, the localization of the B-ABE processed samples was closer to 3D wideband samples (DMOS 4.34), than 3D narrowband samples (DMOS 4.08). However, the difference in DMOS scores is not statistically significant, which can be seen from the overlapping 95% confidence intervals. The average DMOS score for 3D narrowband processing compared to 3D wideband processing was 3.89. This score is statistically different from the 3D WB vs. B-ABE score.

Similar analysis was conducted to the data obtained from the test B. The results are given in the right panel of Fig. 3. In listeners' opinion, the localization of the B-ABE processed samples was closer to 3D wideband samples (DMOS

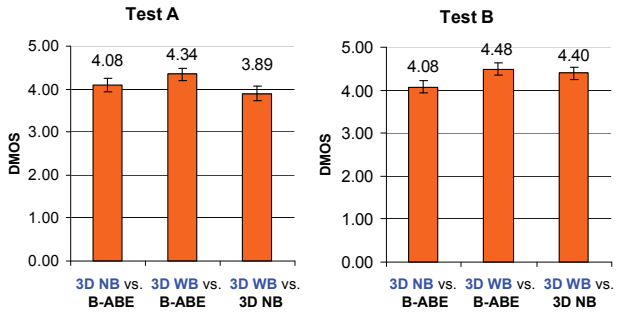

Fig. 3. Results of listening test A (left panel) and B (right panel). The bar heights denote the average DMOS score with 95% confidence intervals given to the second processing in reference to the first processing (highlighted with blue).

4.48), than 3D narrowband samples (DMOS 4.08). This difference is statistically significant. This may indicate that because of small differences between the processings, some listeners have paid attention to localization differences caused by different bandwidths. The average DMOS score for 3D narrowband processing compared to 3D wideband processing was 4.40. This score is not statistically different from the other scores.

The results indicate that the differences between the processings in terms of localization are comparatively small and B-ABE algorithm is able to produce naturally extended speech.

5. CONCLUSIONS

Monophonic artificial bandwidth extension methods cannot be applied as such to binaural speech signals due to a possible mismatch of binaural cues in the highband leading to severe binaural artefacts. The proposed B-ABE method estimates the ITD and ILD parameters from the narrowband binaural signal and adjusts the level and time differences in the artificial highband accordingly. Perceptually the level difference is the dominant localization cue in the high frequencies and the influence of the time difference is less significant. An implementation based on a monophonic ABE algorithm applying spectral folding and frequency domain processing to shape the highband, was also evaluated subjectively. Promising results were obtained from the listening tests. The localization information of the narrowband signal is preserved well in B-ABE processing. Furthermore, the localization of B-ABE signals is even closer to that of the 3D wideband signals than 3D narrowband signals, which proves that the algorithm is able to produce naturally extended bandwidth.

Currently the level difference is constant for all frequencies of the highband but a more sophisticated frequency dependent method based on HRTFs could be used. Another improvement for the method could be a more sophisticated logic for deciding which channel is first processed with the monophonic ABE algorithm. In the proposed implementation, the processing is always applied to the left signal, but possibly some benefit could be obtained by selecting the strongest channel for monophonic ABE processing.

The proposed implementation of B-ABE is based on frequency domain processing but a time-domain implementation could be feasible as well. However, in a time domain implementation, different group delays in cascade signal paths have to be compensated with delay equalization filters, such as allpass filters.

Room effect can improve externalization of sound sources in binaural listening. In the presented listening tests dry input speech signals were used and an artificial room effect was not included in the spatial processing. A future work is to conduct listening experiment to evaluate performance of B-ABE with binaural speech signals including room effect.

6. REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech with one and two ears," *Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, September 1953.

[2] R. H. Gilkey and T. R. Anderson, Eds., *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, New Jersey, 1997.

[3] G. Fant, *Acoustic theory of speech production*, Mouton, the Hague, Netherlands, 1960.

[4] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, 2003.

[5] H. Gustafsson, U. A. Lindgren, and I. Claesson, "Low-complexity feature-mapped speech bandwidth extension," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 2, pp. 577–588, 2006.

[6] 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, Lang. Process.*, vol. 16, pp. 1124–1137, August 2008.

[7] *Multi-lingual speech database for telephonometry 1994*, NTT Adv. Technol. Corp. Available: http://www.ntt-at.com/products_e/speech/index.html, 1994.

[8] O. Kirkeby, E. Seppälä, A. Kärkkäinen, L. Kärkkäinen, and T. Huttunen, "Some effects of the torso on head-related transfer functions," *Presented at the AES 122nd Convention, Vienna, Austria*, May 2007.