Enhancing 3D Audio Using Blind Bandwidth Extension

(PREPRINT)

Tim Habigt, Marko Đurković, Martin Rothbucher, and Klaus Diepold

Institute for Data Processing, Technische Universität München, 80290 München, Germany

Correspondence should be addressed to Tim Habigt (`tim@tum.de`)

ABSTRACT

Blind bandwidth extension techniques are used to recreate high frequency bands of a narrowband audio signal. These methods allow to increase the perceived quality of signals that are transmitted via a narrow frequency band as in telephone or radio communication systems. We evaluate the use of blind bandwidth extension methods in 3D audio applications, where high frequency components are necessary to create an impression of elevated sound sources.

1. INTRODUCTION

An emerging feature in communication systems is 3D audio using Head-Related Transfer Functions (HRTFs). In this scenario, it is desired to place virtual sound sources at different azimuths and elevations to create an immersive three-dimensional impression.

The human auditory system uses different cues to localize sounds. These cues include interaural time differences (ITD), interaural level differences (ILD) and spectral cues [2]. Whereas lateral localization can be achieved using ILD and ITD only, spectral cues are needed to spatialize virtual sound sources at different elevations. Several studies have shown that spectral cues in high frequency bands are needed to perceive the elevation of sound sources. Algazi et al. [1], for example, were able to show that localization accuracy drops significantly if the audio bandwidth is reduced from 22 kHz to 3 kHz. Although elevation cues exist in low frequency regions, high frequency bands play an important role in the perception of elevated sound sources.

Figure 1 shows magnitude plots of HRTFs of a KEMAR dummy-head in the elevation range from 0° to 50°. Elevation-dependent frequency cues can be seen especially in the frequency range above 5 kHz.

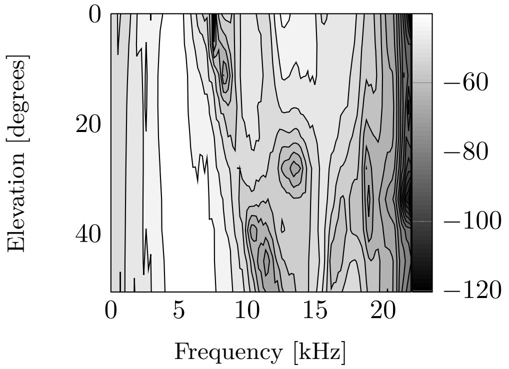

Fig. 1: HRTF magnitude plots of a KEMAR over a range of different elevations. The colorbar indicates the magnitude in dB.

Telephone networks and radio communication systems are generally band-limited. Such systems work, for example, with sampling rates of 8 kHz and therefore only transmit a usable bandwidth of 4 kHz. This degrades the speech quality but the loss of fidelity is usually tolerable as it does not severely impact intelligibility of the communication. On the other hand, this degrades 3D audio quality in band-limited communication systems as the narrowband speech signal does not excite the required high frequency components.

Recently, bandwidth extension or spectral band replication is used in audio codecs like AAC+ [5]. These techniques allow to replicate high frequency spectral features from low frequency information. Therefore, only a small part of the speech bandwidth (eg. 0-5 kHz) needs to be coded and transmitted. The high frequency components (5-10 kHz) can be replicated requiring only little guiding information. Blind bandwidth extension (BWE) recreates high frequency components without any additional information. In previous work, blind bandwidth extension techniques are primarily used to improve the perceived audio quality.

In this work, we evaluate the possibility to create high frequency components using blind bandwidth extension techniques to excite the needed spectral cues to provide an impression of sound source elevation.

Several proposed algorithms for high quality audio bandwidth extension exist. Our main goal is to excite high frequency components of the HRTF to provide elevation cues. We concentrate on accurate localization and do not evaluate the perceived speech quality.

We conducted listening experiments to evaluate this approach. In the listening tests the participants had to judge the elevation of a virtual sound source. Three different sound signals were used that represent the cases of narrowband and wideband speech and broadband noise.

2. BLIND BANDWIDTH EXTENSION

We want to spatialize signals with a bandwidth of 8 kHz and employ a BWE algorithm to extend the bandwidth to 16 kHz. There are several blind bandwidth extension algorithms that aim for high-quality speech reconstruction ([6], [4], [8], [3]). Blind bandwidth extension is generally composed of two tasks. First, the high frequency band of the signal has to be reconstructed. Afterwards, these components are shaped by a filter to match the spectral envelope of natural speech. We employ a bandwidth extension method with low computational complexity proposed by Yasukawa [11]. This processing scheme is composed of several steps that are illustrated in Figure 2.

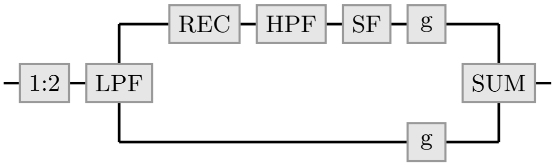

Fig. 2: Blind bandwidth extension scheme.

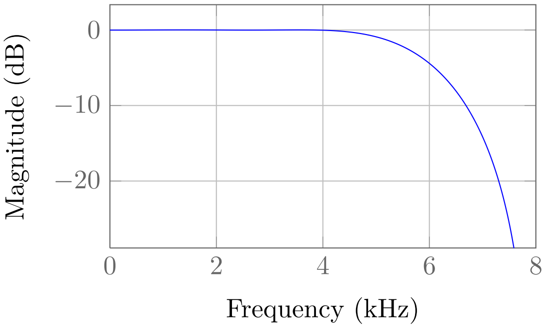

Fig. 3: Transfer function of the shaping filter (SF).

The first step doubles the sampling rate of the signal. The input signal is interpolated by a factor of two by inserting zeros in and applying a low-pass filter (LPF) to remove aliasing artifacts. In the next step the high-frequency components are generated by a non-linear processing element. Yasukawa proposes a full-wave rectifier (REC) as the non-linear filter. This rectifier creates all even harmonics of the input signal. If the input signal is, for example, a sine wave, one can see the harmonics in the Fourier series of

$$|sin(\omega t)| = \frac{2}{\pi} - \frac{4}{\pi} \sum_{n=2,4,6,\dots}^{\infty} \frac{cos(n\omega t)}{n^2 - 1} \qquad (1)$$

in the even multiples of the fundamental frequency ω of the cosine.

The following high-pass filter (HPF) was designed to remove the signal components in the frequencies from 0-4 kHz. This filter removes the DC-offset that is created by the rectification and ensures that the original signal is not distorted. A shaping filter (SF) modifies the spectrum of the signal to better match human voice. The transfer function of the shaping filter is shown in Figure 3. We designed the shaping filter according to the proposed version of Yasukawa. At the end of the processing chain, the generated high frequency components are added to the origi-

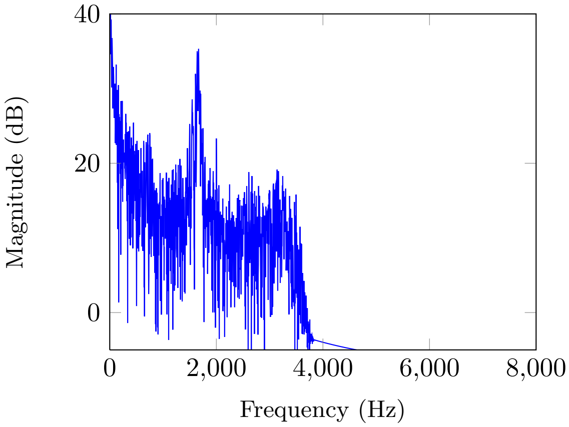

Fig. 4: Spectrum of a narrowband speech signal.

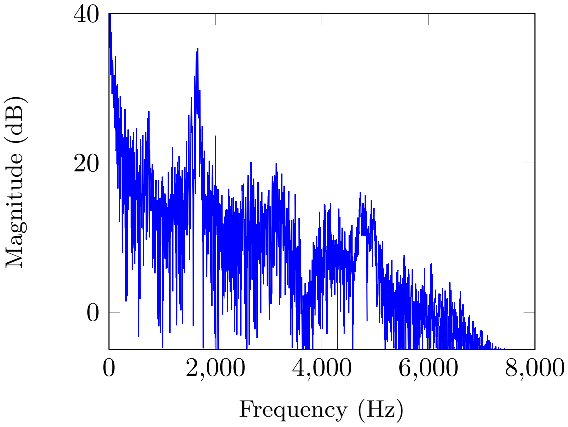

Fig. 5: Spectrum a speech signal after bandwidth extension.

nal low bandwidth signal. Two amplifiers (g) allow to adjust the gain of both paths.
The effect of the bandwidth extension is illustrated in Figure 5 where the spectrum of the narrowband speech signal from Figure 4 is shown after the extension. It can be seen that the original frequency components in the range from 0-4 kHz are not affected by the algorithm.

3. SUBJECTIVE EVALUATION

Algazi et al. [1] were able to show that elevation localization depends on the bandwidth of the signal. They observed that the localization accuracy drops significantly if the audio bandwidth is reduced from 22 kHz to 3 kHz. The reduction of the localization performance is highest in the median plane. For this reason, we chose the positions of the virtual sound sources as follows. The virtual sound sources were positioned on three different azimuths (135°, 180°, 215°), where 0° is directly in front and 90° is to the left of the listener. All the positions are in the back hemisphere to avoid front-back confusions. The elevations were chosen from 0° to 50° with a spacing of 10°. In our coordinate system, 0° denotes the horizontal plane and 90° is directly above. No negative elevation angles were chosen to prevent up-down confusions. We spatialized sound sources by convolving them with Head-Related Impulse Responses (HRIRs) from the publicly available MIT database of a KEMAR [7]. We were not able to measure the individual HRTFs of the test subjects. For this reason, we used HRTFs that were measured with a dummy head that matches the average human torso and head dimensions. The choice of a general set of HRTFs can lead to difficulties as it is not guaranteed that the spatial cues match the test subject. We therefore included broadband noise signals in our sound samples to check if the subjects were able to correctly localize this noise. We used these noise signals as an indicator that the subject could extract correct elevation cues from the HRTFs. The speech sources were taken from VoxForge [10], a free speech database under the GPL license.

3.1. Stimuli

The sound samples can be categorized into three classes:

1. Narrowband speech signals (4 kHz bandwidth)
2. Bandwidth-extended speech signals (8 kHz bandwidth)
3. Wideband noise signals (22 kHz bandwidth)

The speech samples have a duration of about 5 s. The wideband noise signal acts as a reference.

10 subjects aged 22-29 volunteered in the study. The datasets of two participants were discarded as they were not able to correctly localize the broadband noise signal for any spatial configuration and it was assumed that the generic HRTFs did not match their physical characteristics.

3.2. Evaluation procedure

The participants in this study had to decide which one of two sound samples was located at a higher position. Therefore, a randomly selected sound sample was chosen from the database and spatialized at two different elevations on the same azimuth. These virtual sound sources were then presented to the listener with a short pause in-between. Afterwards, the participants had to choose which sample was located higher and mark their choice in a graphical user interface. This procedure is similar to the one used by Kulkarni and Colburn [9].

The graphical user interface as well as the convolution of the sound signals with the HRTFs was implemented in Matlab. The created signals were presented with a pair of Sennheiser HD 380 Pro headphones. The experiments were conducted in a distraction-free environment.

At the beginning of the evaluation, each participant listened to 20 different sound samples from different directions to become familiar with the task and the signals. These results were not included in the evaluation. Furthermore, the participants were informed about the possible locations of the sound sources.

4. RESULTS

Our main goal was to evaluate the impact of blind bandwidth extension on the elevation perception. Figure 6 shows the results of this evaluation.

Three graphs represent the different stimuli. These graphs show the percentage of correct localizations over the angular distance. The smallest elevation distance between two points is 10° due to the HRTF database's spatial sampling grid. The distances 30°, 40° and 50° were combined to get roughly the same number of samples in each of the three elevation distances. All results are shown with their respective 95% confidence intervals.

A localization success rate of 50% corresponds to guessing and indicates that the test subject could not extract the elevation cues necessary for correct localization.

In the case of wideband noise, the participants were able to correctly identify the higher source in 83% of the cases for distances > 30°. This shows that the participants were successfully able to extract the elevation cues although non-individualized HRTFs were used. The percentage of correct localizations for the narrowband speech signal is 50% in this case, which shows that the participants were just able to guess and could not get any information about the elevation.

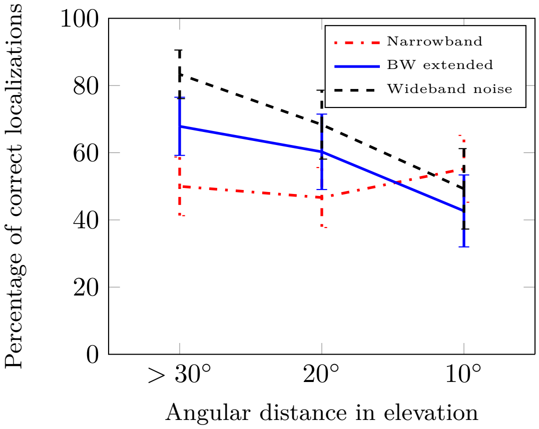

Fig. 6: Results of the subjective evaluation. The plots show the percentage of correct localizations over the elevation distance. Three different stimuli are shown separately. The errorbars represent the 95% confidence interval.

In the case of bandwidth-extended speech, the elevation localization success rate is 68%. This is, considering the chosen confidence level, a significant improvement compared to the narrowband case.

For small distances between the sound sources (10° in this test), the localization success rate lies within the bounds of chance performance.

It was reported that localization in the vertical direction is impaired especially in the median plane. For this reason, we divided the test samples into two groups of azimuths. Half the samples are located in the median plane whereas the other half is at azimuth angles 135° and 215°. In Table 1 the results of this comparison are shown. It can be seen that the elevation localization success rate is roughly 10 percentage points better in positions outside of the median plane for the case of bandwidth-extended stimuli.

5. CONCLUSION

We were able to show that blind bandwidth extension is a valuable tool to improve the reproduction of 3D audio. The bandwidth of speech signals was widened from 4 kHz to 8 kHz using a blind bandwidth extension scheme proposed by Yasukawa.

Azimuth	Wideband noise	Bandwidth extended speech	Narrowband speech
180°	83.9 ± 9.2%	62.0 ± 13.5%	39.2 ± 13.4%
135°, 215°	82.5 ± 11.8%	72.6 ± 11.1%	57.3 ± 11.2%

Table 1: Localization success rate with respect to azimuth.

The subjective evaluation showed a significant improvement of the localization accuracy using generic HRTFs.

6. REFERENCES

[1] V. Algazi, C. Avendano, and R. Duda. Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109:1110, 2001.

[2] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization.* The MIT Press, 1997.

[3] J. Cabral and L. Oliveira. Pitch-synchronous time-scaling for high-frequency excitation regeneration. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[4] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter. Speech enhancement via frequency bandwidth extension using line spectral frequencies. In *IEEE International Conference on Acoustics Speech and Signal Processing*, volume 1. IEEE, 2001.

[5] M. Dietz, L. Liljeryd, K. Kjorling, and O. Kunz. Spectral Band Replication, a novel approach in audio coding. *PREPRINTS-AUDIO ENGINEERING SOCIETY*, 2002.

[6] J. Fuemmeler, R. Hardie, and W. Gardner. Techniques for the regeneration of wideband speech from narrowband speech. *EURASIP Journal on Applied Signal Processing*, 2001(1):274, 2001.

[7] B. Gardner and K. Martin. HRTF measurements of a KEMAR dummy-head microphone. *MIT Media Lab Perceptual Computing*, 1994.

[8] U. Kornagel. Spectral widening of the excitation signal for telephone-band speech enhancement. In *Proc. International Workshop on Acoustic Echo and Noise Control*, pages 215–218.

[9] A. Kulkarni and H. Colburn. Role of spectral detail in sound-source localization. *Nature*, 396(6713):747–749, 1998.

[10] VoxForge. Free speech recognition. http://www.voxforge.org/, Last accessed on 2010-08-30.

[11] H. Yasukawa. Signal restoration of broad band speech using nonlinear processing. In *Proc. European Signal Processing Conference (EUSIPCO-96)*, 1996.