Problem of far-end user's voice in binaural telephony

Tapio Lokki[1], *Heli Nironen*[1], *Sampo Vesa*[1], *Lauri Savioja*[1], *and Aki Härmä*[2]

Helsinki University of Technology
[1]Telecommunications Software and Multimedia Laboratory
[2]Laboratory of Acoustics and Audio Signal Processing
Tapio.Lokki@hut.fi

Abstract

A binaural telephone is a voice-over-ip application which enables auditory telepresence. The basic concepts and implementation issues of such a telephone are discussed. In particular, we will address a specific problem of the voice of a far-end user. In a simple implementation of a binaural telephone far-end user's voice would be located inside the head of the near-end user. We present a solution how this inside-head localization can be reduced automatically in binaural telephony. Finally, we discuss the performance of the presented algorithm and sound quality of the implemented binaural telephone.

1. Introduction

The idea of auditory telepresence is seventy years old. Fletcher [1] report a situation where binaural hearing was examined by letting test persons listen to sounds that were captured at the same time in another room with a dummy head. In this paper we present a two-way telepresence application, called binaural telephone, which is based on the Wearable Augmented Reality Audio (WARA) framework [2]. The presented phone is a voice-over-ip (VoIP) application in which binaural signals are transmitted via a computer network. For sound capturing and playback both users are wearing microphone-earphone transducers, one of which is depicted in Fig. 1. Such transducers enable creation of augmented auditory environment, in which virtual or recorded auditory events can be superimposed to the user's current auditory environment. In addition to voice transmission, the binaural telephone transmits the whole auditory environment to the receiving end. This is a nice feature, e.g., if a near-end user wants to virtually participate in a meeting organized in a far end.

Consider a situation where a far-end user is wearing a microphone-earphone headset and listening the conversation around her/him (see Fig. 2). The near-end user hears the conversation where speakers are distributed in a space as in a real situation. The illusion of telepresence collapses when the local person in the far-end starts to speak, because her/his voice is much louder than other voices. In addition, the voice is localized inside the near-end users head. This sounds unnatural and for convenient use of the system the far-end local speaker has to be rendered so that the sound seems to come outside the head, e.g., in front of the near-end user as illustrated in Fig. 2. Therefore, it is necessary to automatically detect cases when the far-end user is talking.

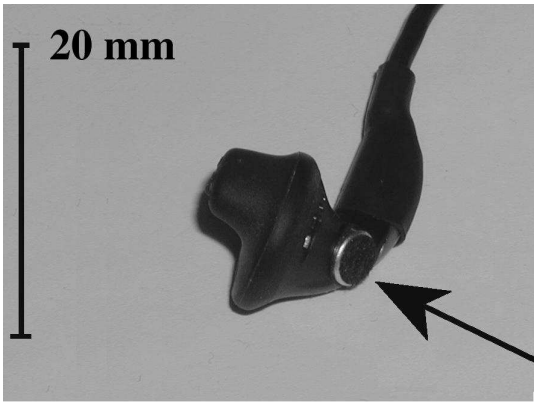

Figure 1: *A two-way transducer with an open-type earphone (Sony MDR-ED268LP). Position of the electret microphone element is indicated by the arrow.*

In the next section we will present a technique with which the voice of far-end user can be detected quite reliably from transmitted binaural signals. Then, in Section 3 we'll present how the segregated speech signal is rendered outside the near-end users head. Finally, evaluation of presented methods is performed and the performance of the system is discussed.

2. Detection of the voice of far-end user

The problem of recognizing the far-end local speaker can be considered to be a voice activity detection (VAD) problem. In this case, there are also other voices present and traditional VAD algorithms cannot be applied. Instead, some sound source segregation method should be utilized. For this problem no ready algorithms were found from literature. So we end up trying several different methods.

The algorithms that were tried were mostly based on

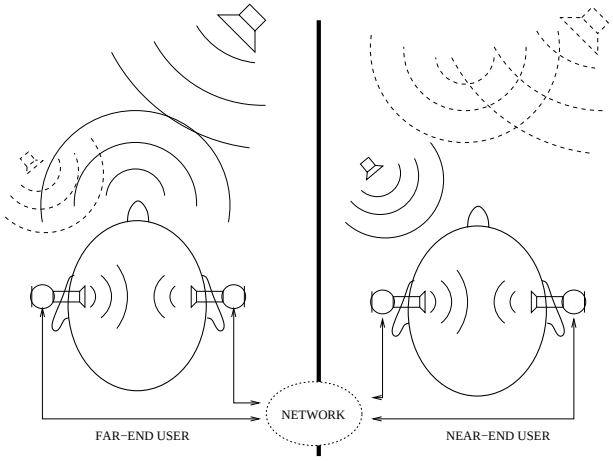

Figure 2: *Binaural telephone. The far-end user's voice should be rendered outside the near-end user's head.*

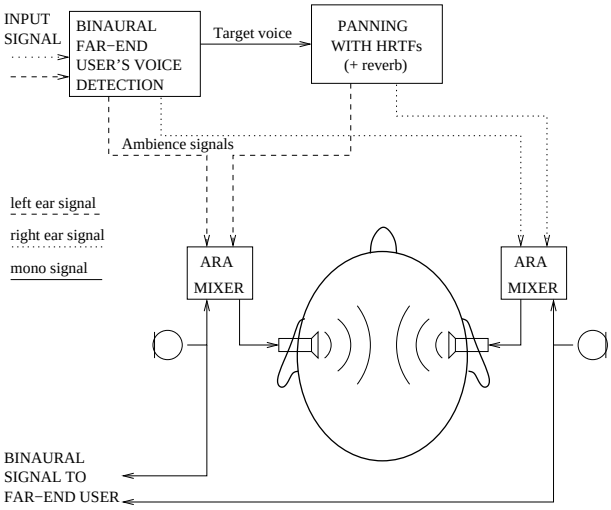

Figure 3: *Rendering of binaural signals in the near end.*

binaural cues. The most important ones being the algorithm described in [3] and the Duet algorithm described in [4] and evaluated in [5]. However, these methods were turned out to be unsuitables for our application, because neither of these methods were found to perform adequately in real reverberant conditions.

The most suitable method for our purposes was finally implemented by modifying a sound segregation algorithm for reverberant conditions [6]. The purpose of the original algorithm was to separate two simultaneous speakers from each other by analyzing signals of two microphones. Here, we have applied same idea which is based on the directional processing of binaural signals. The applied algorithm is called binaural voice activity detector (BVAD).

Figure 3 shows in more detail the signal routes in the near-end. Binaural signals are transmitted with UDP protocol from the far-end and then analyzed in real time with BVAD. Based on analysis results incoming signals are divided into three signals. The detected far-end user's voice (Target voice) is panned with HRTFs in front of the near-end user as depicted in Fig. 2. The detailed signal flow of the applied BVAD algorithm is illustrated in Fig. 4 and is implemented as follows.

The BVAD algorithm assumes that far-end local speaker is in the median plane between two microphones. This is obvious since microphones are mounted in the far-end user's earplugs. All other sounds coming from other directions to the far-end user's ears are called interference signals, as illustrated in Fig. 4. The whole processing starts by dividing continuous signals into 32 ms long time frames. Consecutive frames have 50% overlap. Each frame is then converted to the frequency domain where the rest of the analysis is performed. After FFT, a cross-correlation is computed for each frame pair to analyze the incoming direction of sound. From this correlation the lag of the maximum peak is searched and if it lies within time window that corresponds to less than a defined deviation, e.g. 10°, from the median plane it is judged that the dominant voice is the voice of the far-end local speaker, i.e. Target voice. In other words cross-correlation analysis gives an estimate of the inter-aural time difference (ITD) between the far-end user's ears. In cases where ITD estimation corresponds to more than defined deviation from the median plane the current frame is assumed to represent interference sound. Next, estimates of target voice and interference spectra are computed with a method which is defined by the lag of the cross-correlation maximum. At this point the two obtained signals are an initial estimate of the target voice (Xi) and a crude estimate of the total interference (Yi). Next step in the algorithm is to compare the level of each frequency bin of Xi and Yi signals. In these frequencies where the target voice level is weaker than the interference level the target sound component is set to be zero. This means that far-end local speaker is probably silent. In the opposite case, the interference signal is subtracted from the target sound. Finally, the time domain signal of the target voice is reconstructed with overlap and add using IFFT.

With the presented algorithm the transfered two binaural signals are divided into three signals in the near-end, as illustrated in Fig. 3. Two of them are binaural left and right ear signals (ambience signals) from which the possible far-end user's voice is segregated (has been removed). Naturally, the third signal is the segregated one.

3. Target sound detection and rendering the voice of far-end user

The original algorithm [6], from which our algorithm is adapted, was used as a front end to a speech recognition system or as an aid for hearing-impaired. The segrega-

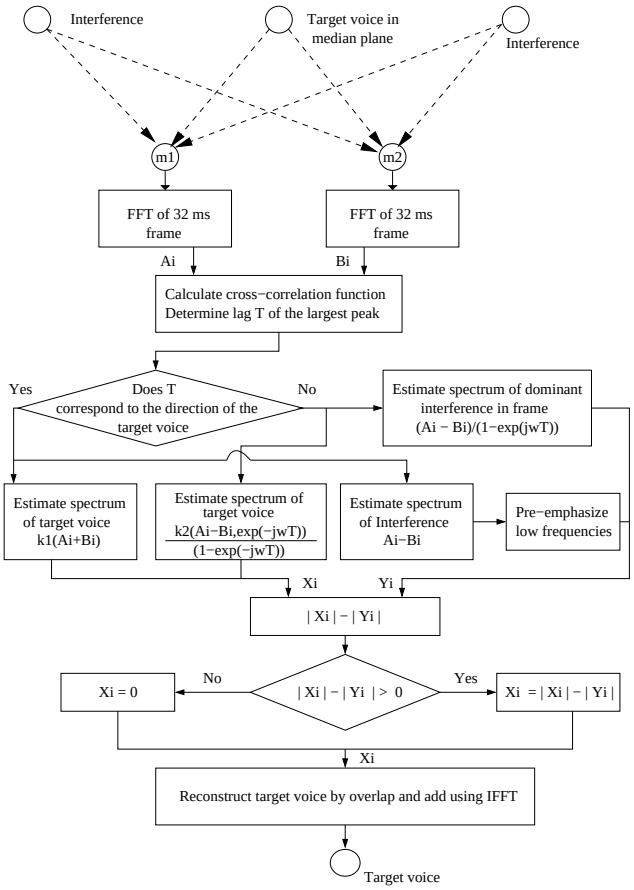

Figure 4: *The applied BVAD and local speaker segregation algorithm for far-end local speaker detection. The algorithm is a modification of sound segregation algorithm [6].*

tion result was not meant to human listeners and so the quality of segregation result was not that critical. This causes some problems when the algorithm is applied to our binaural telephone.

When the target voice region in the median plane is narrow, voice activity detector makes only a few errors, i.e. detects ambience signals as the target voice. However, at the same time it misses the target voice quite often. As a consequence the quality of the segregated sound gets worse. The reason for this is that when the voice activity detector misses the target voice and classifies target as ambience signals it tries to block the target wrongly out. The segregation module switches constantly between the two different sound segregation modes while the far-end local speaker is speaking. When the segregation result is now panned with HRTFs outside the near-end listener's head and mixed in ARA mixer with the ambience signals and the local environment, the far-end local speaker's voice is heard alternately outside and inside the near-end listener's head. This sounds disturbing.

On the other hand, when the region of target voice is increased voice activity detector misses less and less target frames but at the same time the number of detection errors increases. Nonetheless, the quality of the segregated sound gets now better. If the region of the target voice is chosen great enough the result sounds tolerable also after rendering. As a drawback, wrongly detected ambience signal frames cause noticeable disturbance to the final result when they are panned outside the near-end listener's head.

In addition to possible detection errors, rendering of the far-end local speaker's voice can be problematic. When segregated target sound is panned and mixed in ARA mixer, the given location of panned sound might be colored and shifted a bit from a desired position. Shifting occurs because segregation module segregates only the direct sound. The reflections, which belong to the far-end local speaker's voice, are left to the ambience signals. At the same time reflections related to the rendered segregated sound doesn't exist. For these reasons, the rendered voice sounds also colored. Furthermore, it is important to notice that if the levels of separated sound and ambience signals are adjusted to the same or the level of the separated sound is set smaller than the ambience signals the panning result is poor. The reason for this is that now the reflections related to the segregated sound which were left to the ambience signals are louder than the direct sound and might be interpreted as the direct sound. The level of the far-end local speaker should though be set clearly higher than the level of the ambience signals. The louder the far-end local speaker's voice is, the better this voice is heard as coming outside the near-end local speaker's head after rendering. The drawback is that at the same time the sound gets more colored. The best result has been achieved by adding some diffuse reverberation to the separated sound, since reverberation helps in externalization and improves the impact of panning.

4. Evaluation of the system performance

The performance of sound segregation algorithm's voice activity detector was evaluated. This was done by specifying the percentages of frames classified wrongly as target when they should have been classified as ambience or surround signals, and vice versa. For evaluation, two recordings were done at a typical office environment. In the first one, an additional speaker, considered as ambience signals, was speaking in 90 degrees azimuth at the distance of one meter related to the local speaker. In the second case, additional speaker was located at 45° angle, one meter from the local speaker. The percentage of wrongly detected frames (errors) and missed frames are presented in Tables 1 and 2.

The results show that with small angles -10° - 10° the segregation result is poor because almost half of the frames which should have been classified to belong to the local speaker are classified wrongly. Informal listen-

Azimuth	Errors: (%)	Missed: (%)
-80° - 80°	25	22
-70° - 70°	23	22
-60° - 60°	22	23
-50° - 50°	21	24
-40° - 40°	18	25
-30° - 30°	13	29
-20° - 20°	8	35
-10° - 10°	3	43

Table 1: *Detection errors in case of two simultaneous speakers. An additional speaker at 90° azimuth and at 1 meter distance.*

Azimuth	Errors: (%)	Missed: (%)
-40° - 40°	20	21
-30° - 30°	15	27
-20° - 20°	8	35
-10° - 10°	3	43

Table 2: *Detection errors in case of two simultaneous speakers. An additional speaker at 45° azimuth and at 1 meter distance.*

ing test also supports this result. By listening the recordings and interactively adjusting parameters it was found that a tolerable quality is obtained when less than 30% of frames are missed.

One interesting question is, whether the voice activity detector is needed at all if its performance is not better than what was presented above. By listening, it can be found that the bubbling caused by misdetected frames does not sound so disturbing when compared to the case where the module switches frequently between two different sound segregation modes. By defining that all the time target voice region is separated from ambience signals the computational load caused by the algorithm could be reduced.

5. Discussion

The presented system sounds quite reliable, although the performance is not perfect in the sense that far-end local speaker is not completely segregated from other sounds. The complete separation would be a hard problem to solve and more advanced methods to the sound segregation and voice activity detection are needed.

If far-end user's voice is needed to keep in stable position, a head-tracker device is required. The head-tracker provides orientation and location of near-end user's head and this information can be utilized in panning. Of course binaural telephone can be used without head-tracking, but more convenient telepresence is obtained with a tracked rendering.

6. Conclusion

Binaural telephone enables a certain degree of telepresence, because the whole 3-D auditory environment is transmitted to the near end. As explained in this paper the voice of the far-end user can make transmitted auditory environment unnatural. To solve this problem we have presented a method with which the telepresence experience can be enhanced. In addition, the performance of the presented algorithm is evaluated and it has been found that it works quite well, but not perfectly. Thus, the sound quality of the implemented binaural telephone is good enough for demonstration purposes. In the future, to have better sound quality, more advanced techniques should be applied for each individual component of the system.

7. Acknowledgments

This work has been carried out in collaboration between Helsinki University of Technology and Nokia Research Center.

8. References

[1] H. Fletcher, "An acoustic illusion telephonically achieved," *Bell Laboratories Record*, vol. 11, no. 10, pp. 286–289, June 1933.

[2] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, H. Nironen, and S. Vesa, "Techniques and applications for wearable augmented reality audio," in *the 114th Audio Engineering Society (AES) Convention*, Amsterdam, the Netherlands, March 22-25 2003.

[3] N. Roman, D. Wang, and G. J. Brown, "Speech Segregation Based on Sound Localization," Department of Computer and Information Science, The Ohio State University, Columbus, OH, Tech. Rep., 2002, Technical Report OSU-CISRC-6/02-TR16.

[4] S. Rickard and Ö. Yilmaz, "On The Approximate W-Disjoint Orthogonality of Speech," in *Proceedings of ICASSP2002*, vol. 1, Orlando, FL, USA, May 2002, pp. 592–523.

[5] M. Baeck and U. Zölzer, "Performance Analysis of a Source Separation Algorithm," in *Proceedings of the 5th International Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, September 2002, pp. 207–210.

[6] A. Shamsoddini and P. Denbigh, "A sound segregation algorithm for reverberant conditions," *Speech Communication*, vol. 33, pp. 179–196, 2001.