BINAURAL POSITIONING SYSTEM FOR WEARABLE AUGMENTED REALITY AUDIO

Miikka Tikander, Aki Härmä, and Matti Karjalainen

Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
P.O.Box 3000, FIN-02015 HUT, Finland
miikka.tikander@hut.fi
aki.harma@hut.fi
matti.karjalainen@hut.fi

ABSTRACT

Microphone arrays have been traditionally used for locating sound sources. In this scenario the array is assumed to be static and the sound sources may be static or moving. If, on the other hand, the sound source or sources are assumed to be static and a user is wearing a microphone array the exact position and orientation of the user can be solved. In this paper we propose a method for binaural positioning of a user based on binaural microphones and known external anchor sources placed in the environment of a user. We demonstrate that user's location and orientation can be tracked in many cases very accurately even if the anchor sound is at a low level and there are distracting sources in the environment.

1. INTRODUCTION

In wearable augmented reality audio (WARA) systems the acoustic environment around a user is enriched by adding virtual audio objects to the environment. Typically, the objective is to produce an impression that virtual sources cannot be discriminated from real sources around the user. One potential way of achieving this is based on a system illustrated in Fig. 1 [1]. Here, the user is wearing a specific headset where a pair of microphone elements has been integrated with in-ear headphones. When signals from the binaural microphones are directly routed to the headphones the user is exposed to a slightly modified representation of the real acoustic environment. The addition of virtual sources takes place in the augmented reality audio (ARA) mixer in Fig. 1.

Preliminary listening tests [1] indicate that this system is very promising for production of realistic augmented reality audio experience for a user. However, rendering of virtual sources in such a way that they have a static location in the augmented environment requires tracking the position and orientation of the users head. In this article we explore the possibility to track these using the binaural microphone signals which are already available in the system (See Fig. 4) thus avoiding the need for additional hardware for tracking.

Acoustic head-trackers are commercially available but most of them are based on a configuration where user is wearing an *user element* emitting ultrasound and head-tracking based on processing of signals from a microphone array placed in the environment. Sometimes this is called an *outside-in* system for tracking [2]. In this article, we propose an *inside-out* system where user is wearing a binaural microphone array and the source, *anchor*, is placed in the environment. The most obvious reason for this selection

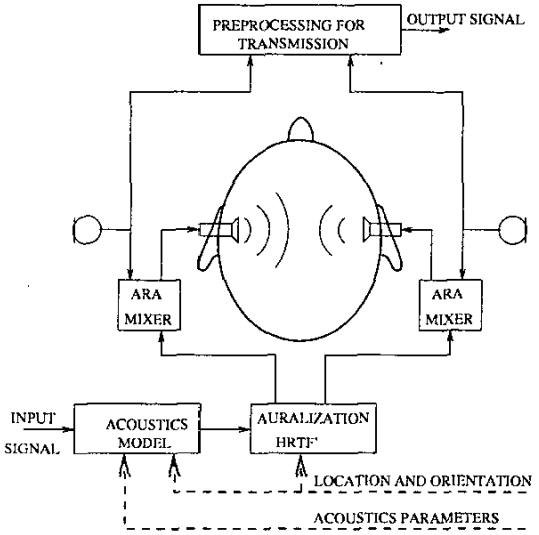

Figure 1: *Wearable augmented reality audio system based on a specific headset and signal processing architecture.*

is that the power consumption of a loudspeaker is much higher than that of a receiver (microphone array) which makes the latter more viable for wearable devices. Secondly, tracking information is needed in the system worn by the user.

In a typical acoustic environment we may identify many potential anchors such as computers or air shafts in an office environment. However, in the current paper we concentrate on a special case where the signal emitted by the anchor is known. One or more sound anchors are emitting known reference signals in the proposed system. When the reference signal is known the rejection to interfering sounds can be greatly improved through correlation between the recorded and emitted signals. With multiple sound sources the reference signal can be divided into non-overlapping frequency regions to increase source separation. The results are shown for an anechoic situation and for a reverberant room condition. The results are also compared with a traditional electromagnetic position tracking device.

0-7803-7850-4/03/$17.00 © 2003 IEEE

2. BINAURAL POSITIONING METHOD

Technologies for source localization with microphone arrays have been reviewed recently, e.g., in [3]. Binaural microphone arrays for source localization have also been presented, e.g., in [4]. In this article we introduce a head-tracking system based on processing of binaural signals.

The time delays of arriving sounds are derived from peak locations of the cross-correlation functions. The generalized cross-correlation methods [5] are very prone to interfering sound sources as the correlation is calculated between the microphone signals. If a source signal is known the correlation can be calculated between the known source signal and the recorded signals. This increases the robustness of localization considerably. With one known source the relative distance and lateral angle can be estimated. With more sources the exact 2D or 3D position of a user can be estimated.

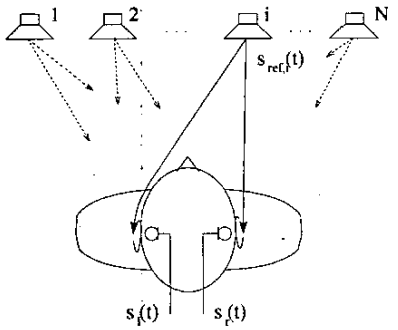

Figure 2: *Schematical view of a system with N static sound sources.*

2.1. Estimating the ITD

Let us assume there are N static sources, anchors, in the system (see Fig. 2). Each anchor is emitting (looping) a known reference sound sample and all the samples are divided into smaller non-overlapping groups of subbands with a filter $P_i(f)$. In the following the procedure of calculating the *interaural time difference* (ITD) for one anchor is introduced. For the other anchors the calculation is identical.

The reference signal emitted by the ith anchor is given by

$$X_{\mathrm{ref},i}(f) = X_{\mathrm{ref}}(f)P_i(f), \tag{1}$$

where $X_{ref}(f)$ is the Fourier transform of the original known reference signal and $P_i(f)$ is the filter for dividing the signal in subbands.

The cross-correlation between the known reference and the recorded signal is given by

$$R_{\mathrm{l,ref},i} = \int_{\tau-T}^{\tau} \Phi_i(f) S_{\mathrm{l,ref},i}(f) e^{j2\pi f\tau} df, \tag{2}$$

where

$$S_{\mathrm{l,ref},i}(f) = E\{X_{\mathrm{l}}(f) X_{\mathrm{ref},i}(f)^*\} \tag{3}$$

is a sampled cross-spectrum between the recorded left channel and the reference signal. The weighting function $\Phi_i(f)$ is given by

$$\Phi_i(f) = \frac{P_i(f)}{|S_{\mathrm{l}}(f) S_{\mathrm{ref},i}(f)|^{\gamma}}, \tag{4}$$

where $P_i(f)$ is a frequency mask for the ith source and is the same as used for filtering the anchor signal. γ is a parameter for changing the amount of magnitude normalization (GCC $\leftrightarrow$ PHAT) [5]. In the same manner, the cross-correlation $R_{\mathrm{r,ref},i}$ is calculated for the right channel as well. Now the estimate for the ITD_i is given by the distance of the maximum values of the $R_{\mathrm{r,ref},i}$ and $R_{\mathrm{l,ref},i}$.

$$\mathrm{ITD}_i = \mathrm{maxarg}(R_{\mathrm{l,ref},i}) - \mathrm{maxarg}(R_{\mathrm{r,ref},i}). \tag{5}$$

When the ITD is known the lateral angle of the user relative to the i:th anchor can be solved. The ITD for the rest of the anchors is solved the same way.

2.2. Estimating the distance

When the distance between an anchor and the microphones changes, as the user moves, the maximum values in the cross-correlation responses move accordingly. This information can be used to estimate the user movement. The chance in distance relative to the ith anchor is the average movement of the maximum values of the cross-correlation responses. The change in distance d_i in samples to the ith anchor can be estimated with

$$d_i = \frac{1}{2}(\mathrm{maxarg}(R_{\mathrm{l,ref},i}) + \mathrm{maxarg}(R_{\mathrm{r,ref},i})) - c, \tag{6}$$

where c is the initial value of d_i.

When the distances to all of the anchors are estimated the angle and position of the user can be estimated via circulation. With two anchors a 2D-position and with three anchors a 3D-position can be estimated. Though, increasing the number of anchors increases the robustness of the system and allows lower pressure levels for the anchors.

2.3. Synchronization

To make anchors perceptually as unnoticeable as possible the sound samples played from the anchors need to be made longer than the frames used in computing. For this reason the system needs synchronization.

In our system, synchronization is done as follows. Assuming the reference signal $s_{\mathrm{ref},i}(t)$ played in the ith anchor is n samples long (See Fig. 2). When the tracking system is started the first n samples are buffered. Then a cross-correlation between the ith reference signal $s_{\mathrm{ref},i}(t)$ and the other channel of the recorded signal $s_{\mathrm{r}}(t)$ or $s_{\mathrm{l}}(t)$ is calculated. Then the reference signal in the memory is circularly shifted to synchronize the system with the recorded signal. After synchronization the recorded frame should match a frame from the reference signal. The same synchronization is repeated for each anchor separately.

As the user moves the synchronization degrades and eventually when the recorded frame and the frame from the reference signal does not correlate enough the system needs to be re-synchronized. This is done the same way as the initial synchronizing. The on-line synchronization can be performed in the background while running the system without interrupting tracking. The on-line synchronizing needs only to be done for the anchors that are running out of synchronization. This reduces the computational load compared to the initial synchronization.

3. RESULTS

The method was tested in anechoic and reverberant conditions. Fig. 3 shows a schematical view of the measurement setup. For

the measurements only one anchor (source 1) was used and source 2 was used for playing interfering sounds. A change in distance is considered as a movement along the y-axis and an angle of zero degrees corresponds to the case when the user is facing source 1 (see Fig. 2). The user was also wearing an electro magnetic (Polhemus ActionTrak) motion tracker receiver which was placed on the top of the user's head.

In the measurement the sampling frequency was 44.1 kHz and the frame size in the calculations was 1024 samples. A 32768 point sequence of white noise was used as a known reference signal.

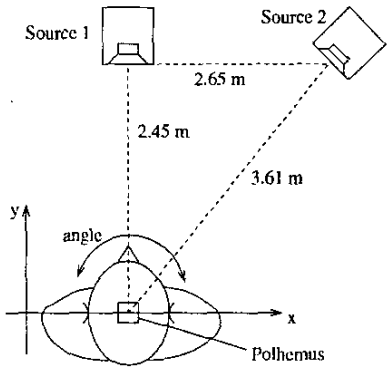

Figure 3: *Schematical view of the measurement setup. Source 1 is used as a reference and source 2 is used as an interfering sound source.*

Figure 4: *Headset used in the measurements. Arrow points at the embedded microphone.*

3.1. Rejection to interfering sounds

For measuring the sensitivity to interfering sounds source 2 in Fig. 3 was used to play steadily increasing white noise. Source 1 was emitting (looping) the reference signal at 37 dB(A). The user was standing in front of the anchor source and turned his head repeatedly to the left and right.

Fig. 5 shows the results of the measurement. Second and third panels (from the top) illustrate cross-correlation responses between the the reference and the recorded signals. The responses are zoomed to show the vicinity of the maximum value of the correlation. From the cross-correlation responses can be seen that at larger angles the head is shadowing the other microphone and the correlation decreases resulting in a decreased accuracy in the estimate of user's position. Especially when the interfering sound level is high and the reference level is low this effect is pronounced. The effect can be seen in the lower two panels where the binaural

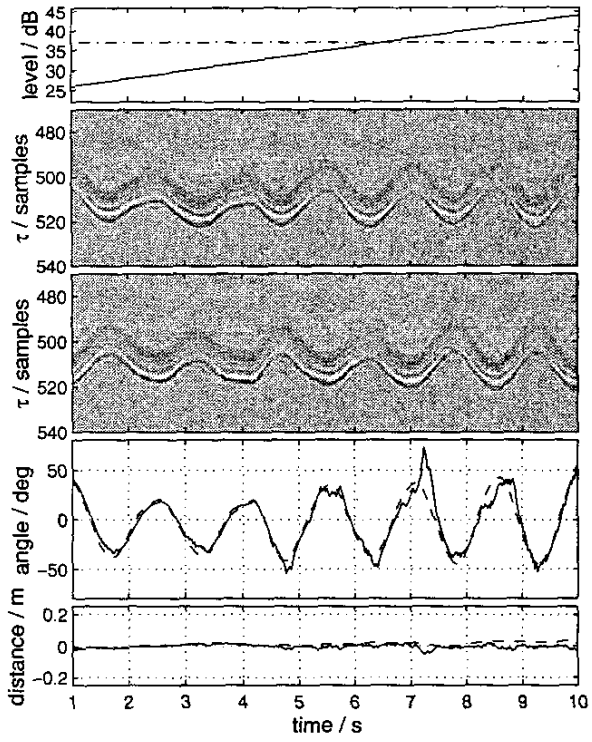

Figure 5: *Top: Solid line is the interfering source level and the straight line is the reference source level (37 dB(A)). Middle and upper middle: Correlation responses at both recorded channels. Lower middle and bottom: Estimated angle and distance. Solid line is the binaural tracking and the dashed line is the data from the electro-magnetic tracker.*

angle and distance estimates (solid line) are plotted. The dashed line is the electro-magnetic positioning device data.

3.2. Reverberant room

Fig. 6 illustrates the results for a measurement performed in a reverberant room (4.5 m x 10.5 m, T_{60} = 0.2 s). The background noise level in the room was 40 dB(A) and the reference signal level was 53 dB(A). During the measurement the user walked back and forth in front of the loudspeaker while turning his head to the left and right. The setup is shown in Fig. 3. Source 1 was used as an anchor and source 2 was not used in this experiment.

The two upper panels show cross-correlation functions between the reference and the recorded signals in the two ears. The two lower panels show estimated angle and distance in the proposed binaural tracking method (solid line) and in an electro-magnetic positioning device (dashed line).

3.3. Positioning in subbands

When multiple anchors are used, source separation can be increased by dividing the reference signals in each anchor into smaller non-overlapping frequency regions. Fig. 7 plots the mean error in the angle estimate when changing the frequency mask P_i. The

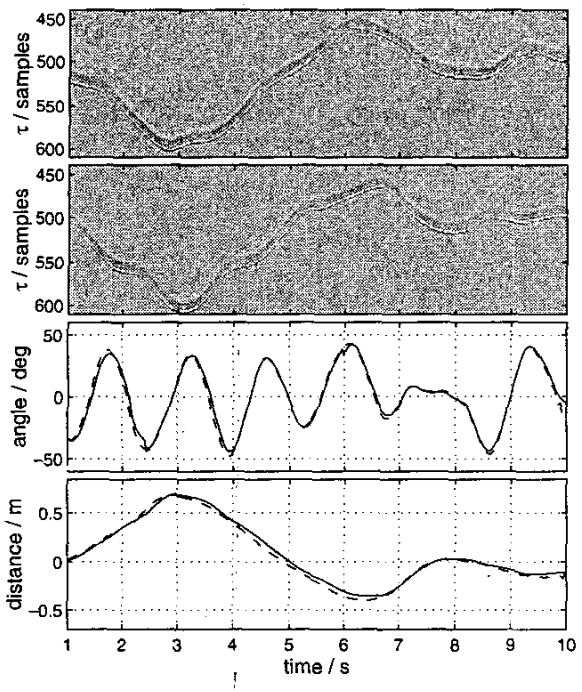

Figure 6: *Measurement in a reverberant room. Background noise level was 40 dB(A) and the reference level was 53 dB(A). Top and upper middle: Correlation responses for both recorded channels. Lower middle and bottom: Estimated angle and distance. Solid line is the binaural tracking and the dashed line is the data from the electro-magnetic tracker.*

upper panel shows the used frequency mask. The bins in the mask were widened for each measurement to see the effect of the P_i. The x-axis in the lower panel tells how many percents of the frequency band were used for the anchor. The data used for the measurement is the same as in Fig. 6.

4. DISCUSSION

In this article we have introduced a method for positioning of a user by binaural microphone signals. The proposed system is based on the use of external audio sources placed in the environment of the user. When the source signal or signals are known the 3D-position and the lateral angle of the user can be estimated. The system was tested in both anechoic and reverberant conditions. Tracking results of the proposed system were compared with an electro-magnetic position device. The results of the measurements were very promising. The system also worked well in reverberant conditions.

The proposed system uses microphones of a specific binaural head-set which is a part of a wearable augmented reality audio terminal. No additional hardware is needed for tracking. It is also possible to use the same binaural signals for localizing unknown sources and estimation of the acoustical properties of the environment.

In this article, wideband white noise was used as anchor signal. It was demonstrated that the anchor sound levels can be kept

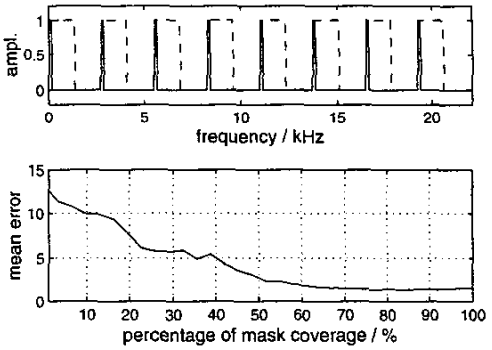

Figure 7: *Upper figure shows the frequency mask P_i used in the measurement. The width of the bins was varied from very thin to a totally flat response. The lower figure plots the mean deviation of the angle estimate error as the mask changes compared to a data from an electro-magnetic positioning device. The measurement data is the same as in Fig. 6*

low even if there are interfering sound sources in the environment. However, hissing noise from the anchor source may be disturbing in some cases. For a user with proposed headset anchor sounds can be easily canceled because they are apriori known. One could also use more pleasant anchor signals, e.g. music, as reference, or hide anchor sounds into music signals utilizing masking properties of hearing.

5. ACKNOWLEDGMENT

This work has been carried out in collaboration between Helsinki University of Technology and Nokia Research Center. A. Härmä's work has been supported by the GETA graduate school and the Academy of Finland.

6. REFERENCES

[1] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, H. Nironen, and S. Vesa, "Techniques and applications of wearable augmented reality audio," in *AES 114th Convention Paper*, Amsterdam, The Netherlands, March 2003.

[2] K. Mayer, H. L. Applewhite, and F. A. Biocca, "A survey of position-trackers," *Presence: Teleoperators and virtual environments*, vol. 1, no. 2, pp. 173–200, 1992.

[3] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds. Springer-Verlag, 2001, ch. 7, pp. 131–154.

[4] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, Hong Kong, May 2003.

[5] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech and Signal Processing*, vol. 24, pp. 320–327, August 1976.