July 2020

ADVANCED HEADSET TRACKING FOR OPTIMIZED USER EXPERIENCE WITH DIRECTIONAL AUDIO

Bjorn Winsvold

Marcus Widmer

Asbjorn Therkelsen

ADVANCED HEADSET TRACKING FOR OPTIMIZED USER EXPERIENCE WITH DIRECTIONAL AUDIO

AUTHORS:

Bjorn Winsvold
Marcus Widmer
Asbjorn Therkelsen

ABSTRACT

A solution is provided that correctly positions a sound source for a headset user in a video call, by first finding the orientation of the head using short ultrasonic bursts and finding the difference in time-of-flight using one microphone at each ear. This head orientation, combined with the location of the participants on the screen, is then used in the binaural processing to render the sound source position correctly.

DETAILED DESCRIPTION

Some collaboration endpoints have support for superior directional audio experience in a multi-party video call by using the dedicated loudspeakers to play out sound from different locations separately. Since this sound output is correlated with the video-layout, this gives a good user experience that makes it easy for the user to follow the conversations. This is implemented such that the left talker will be played out in the left channel of a stereo or three channel signal (L,C,R) and the right talker is played out in the right channel only and this gives excellent directional perception of the different sources. See Figure 1 below, where the circle represents the head of a user.

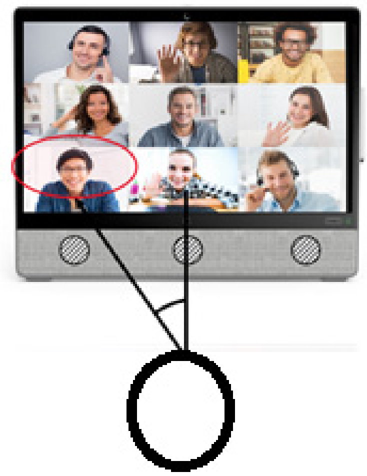

Figure 1

In some cases, the user wants to use a headset for privacy reasons in video calls. When playing out the signal from the leftmost talker to a headset, the signal will be in the left channel only and that will give a very annoying user experience. The user will hear the talker at 90 degrees to the left, but visually the talker is located 20-30 degrees to the left. See Figure 2 below.

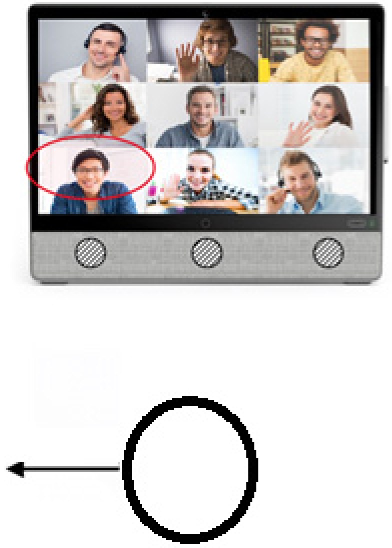

Figure 2

This can be solved by reducing the audible panning span in the headset to for instance -30 to 30 degrees. See Figure 3 below. However, another issue arises. The user hears a person talking at an angle of 20 degrees to the left and he/she wants to put eyes on this person. But when the user moves his head 20 degrees and looks into the person speaking, the sound in his headset has also moved 20 degrees further to the left, and now he hears the person at 40 degrees (relative to the screen center = 0 degrees).

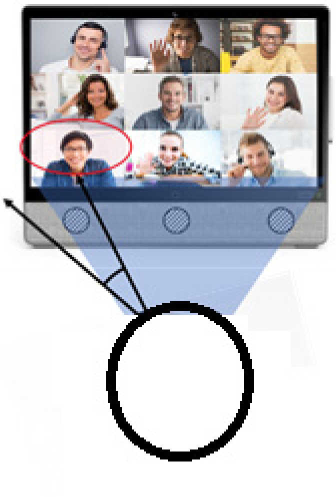

Figure 3

A solution is presented herein that includes a mechanism for estimating the distance and the angle of the head of the user relative to the screen. Head tracking using gyro or accelerometer is known, but by utilizing the ultrasound capabilities of a collaboration endpoint and the two microphones on a headset, the head angle can be estimated in a new way using pre-existing hardware. When a short ultrasonic burst is transmitted from the endpoint, the head-angle ('a' in Figure 4), is determined as a function of the time delay (t) between two microphones (blue and orange plots in Figure 4), ear-to-ear distance (d) and the speed of sound (c).

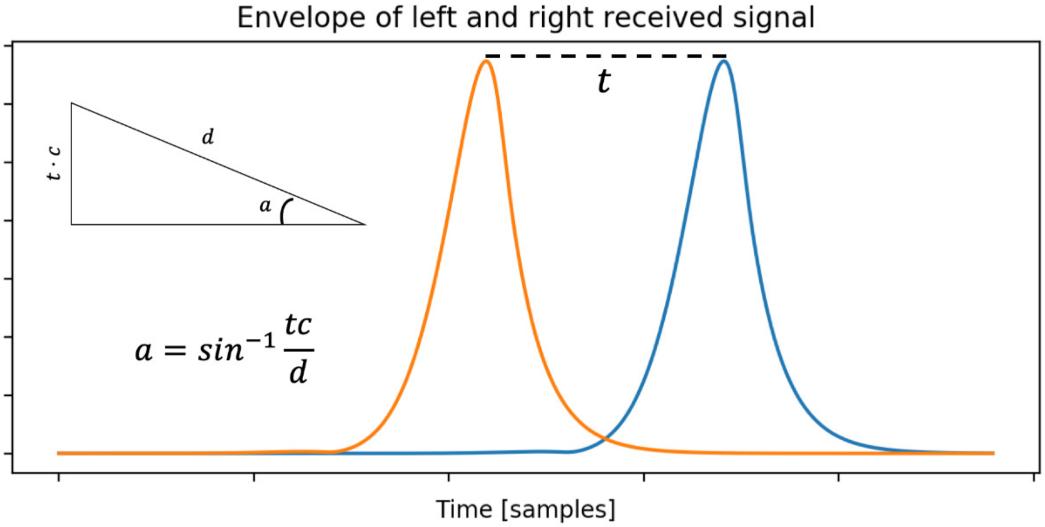

Figure 4

Estimating the distance is important to derive the right perspective when rendering the three-dimensional audio. To do this, time-of-flight measurements are made using the ultrasonic burst. The hardware latency can be determined by evaluating when the noise burst arrives at a microphone on the video endpoint. By observing the midpoint between the left and right signal and subtracting the hardware latency, the distance to the user may be determined.

Information is needed from the endpoint microphones to estimate the distance. Therefore, the audio processing also needs to be performed on the endpoint itself. Because the estimation is based on a signal coming from the endpoint, angle calibration is not needed - as it is for gyro/accelerometer based solutions. Alternatively, the head angle can be estimated by using machine learning combined with camera information, but angular resolution might be a challenge here.

A second aspect of this solution involves using the estimated angle to position the sound source correctly. By using Binaural processing, the audio-origin from each participant will match the video-origin in a conference call on a collaboration endpoint when using headphones. Pre-defined non-user specific head-related-transfer-functions (HRTFs) may be used to do this. Because all human ears are different, the HRTFs should ideally be individually measured, but this is difficult to achieve in practice and the quality one gets from a generalized HRTFs is typically "good enough" for this application.

Traditional HRTF processing assumes that the user sits completely still and does not move his head. This solution aims to make directional audio work for an arbitrary head orientation. The estimated head angle is used to compensate for the head movement when doing the binaural audio processing. In this way, the sound source is fixed in one place.

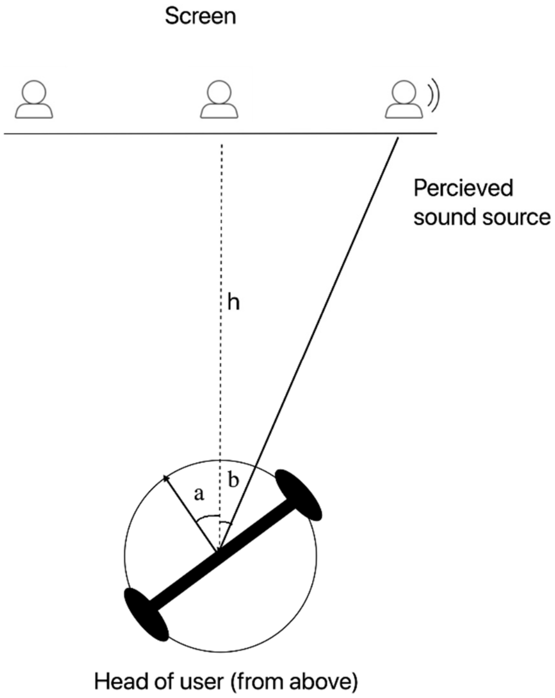

Figure 5

The processing itself starts with a mono source of audio. If a user is in a meeting with 3 other participants, the video endpoint has access to 3 individual audio streams with layout-position-metadata. By knowing the distance ('h' in Figure 5), from the endpoint and the user, the layout-position can be converted to a horizontal (angle 'b' in Figure 5). By subtracting the head angle ('a' in Figure 5) from the horizontal angle (b) the angle between the ears and the sound source is obtained.

The horizontal angle resolution of the used HRTF dataset may be 5 degrees. The obtained angle is then used in a lookup table to find the HRTF corresponding to the closest horizontal. The mono sound source is then convolved with the transfer function for left and right ear, producing two channels of audio. Doing this for all participants achieves three-dimensional rendering of sound sources in the meeting.

In summary, the need for headphones in open work environments is increasing. When using headphones in a call today, the sound seems to be coming from somewhere inside your head. This does not really align with experience of real life face-to-face communication. The sound should be coming from the location where you see and interact with the person. This solution aims to position the sound source correctly by first finding the orientation of the head using short ultrasonic bursts and finding the difference in time-of-flight using one microphone at each ear. This head orientation, combined with the location of the participants on the screen, is then used in the binaural processing to render the sound source position correctly.