Head Tracking Binaural Localization System for Horizontal Sound Source Detection *

Nolan Lem[1] and Jens Ahrens[2]

[1]Center for Computer Research in Music and Acoustics (CCRMA) Stanford University
[2]Technische Universität Berlin, Germany

***Abstract*— This project attempted to reconstruct an interactive spatialization auditory scene by using a head tracking sensor in conjunction with a binaural panning system implemented in the MaxMSP programming environment. This study was motivated by the need to be able to demonstrate in more informal settings–such as classroom demonstrations, or paper presentations–the general experimental environment that many researchers have used to examine the localization of point-like sound sources on the horizontal plane.**

I. INTRODUCTION

Sound localization tasks involving point-sources. are well-studied in the psychoacoustics literature. The human auditory system exploits several auditory features and percepts to infer source location from an incoming sound. These include ITDs, ILDs, spectral cues, head movement cues, intensity/loudness cues, familiarity to the sources, Direct-to-reverb ratios (DRR), and visual and other non-auditory cues (cite). Among these, the last four are associated with distance cues while the others allow us to align our direct localizations. It is well known that have much more spatial accuracy in identifying sound sources on the horizontal plane as compared to our vertical (azimuth) source detection. Head-related Transfer Functions (HRTFs) describe the frequency responses of our ears. They can be characterized by a filter that describes how a sound from a specific source will arrive at each of the ears. Because of the localization techniques we use to detect sound in space, we can simulate "virtual" sources by taking advantageous of these transfer functions typically by deriving head-related impulse responses (HRIRs) which is the Fourier transform of the HRTF. The HRTF reflects the signal differences each ear receives given its orientation in space and the shape of a person's head and ear canal.

When multiple sources are involved, we often create sound-object representations that facilitates our ability to understand and parse out information from an auditory scene. This fusion of synchronous auditory streams (or otherwise) is usually a function of how well temporally correlated the signals are which results in summing localizations from which point-like virtual sources emerge. In relatively reflection-free acoustic spaces, there are several factors that influence the perception of sound sources. These include the signal level or duration which tends to increase the perceived width of the distribution, and the frequency of the source stimuli which tends to decrease the width percept. Typically, studies have used the "concurrent minimum audible angle (CMAA)" to discriminate between listeners' abilities to discern angular separation among sources presented as point-like sources projected from multiple speaker channels arranged in different spatial orientations. Because of the effects of stimulus frequency, researchers have studied a plethora of stimuli containing various spectral content to examine its effects on listeners' ability to detect sound sources in spatial configurations. Additionally, researchers have often employed the use of broadband noise in the context of point-source identification tasks.

nlem@ccrma.stanford.edu
jens.ahrens@tu-berlin.de

II. CONVOLUTION WITH HRIRS

A. Description

This system can be formulated by the following description. Let a group of left and right ear HRIRs, $h_i^l(n)$ and $h_i^r(n)$, each of size N and separated by 1° of angular horizontal separation be convolved with a source signal, $x_{source}(n)$. (1).

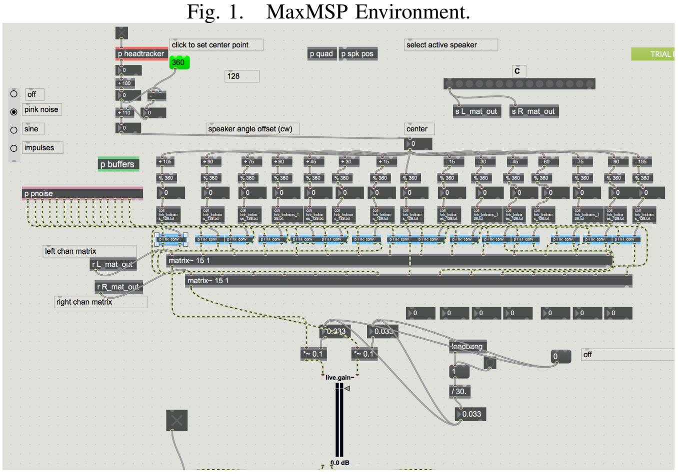
$$\begin{aligned} y_{i,t}^l(n) &= (h_i^l * x)(n) \\ y_{i,t}^r(n) &= (h_i^r * x)(n) \end{aligned} \tag{1}$$

where i indexes HRIR associated with each degree of separation and $y_i^l(n)$ and $y_i^r(n)$ are the outputs of source signal convolved with the respective HRIRs for the left (l) and right(r) ears respectively. Let $W(n)$ be an equal-power window function also of size N that "crossfades" between two $h_i^{l,r}(n)$ impulse responses when we traverse (an auditor moves their head) along two sequential angles, y_i from the y_{i-1}.

$$y_{i,t}^{l,r}(n) = \sum_{t=1}^{T} W(n - \frac{2Nt-1}{2}) y_{i-1}^{l,r}(n - Nt) \tag{2}$$

where $y_{i,t}^{l,r}$ is the output signal of length NT that accumulates to characterize the auditor's head movements from time t to T.

Fig. 1. MaxMSP Environment.



III. Binaural Spatial Localization System

A. System Description

This system creates an array of 15 virtual point-sources that act as separate circular-distributed channels using HRIRs from the HRTF Dataset for Horizontal Localization from Deutsch Telekom Laboratories (2008) in Berlin (Lindau 2008). In this dataset, individual HRIRs were gathered using a binaural "dummy" head from point-sources at 1.0 increments to create 360 separate HRIRs corresponding to different head orientations corresponding to the virtual listener turning their head. For computational reasons imposed by MaxMSP, these 256 sample HRIRs were downsampled to be 128 samples long. The sample buffer object in MaxMSP, *buffir* , was used to convolve the HRIRs with pink noise, sine tones, and impulses. These source sounds were selected because they represent common signals typically administered in psychoacoustics studies in sound localization tasks. We used a 3-axis accelerometer (cite) in conjunction with an arduino Uno connected through the serial port to output the yaw (azimuth position) of the auditor's head position which indexes the HRIR orientation to be convovled with the source sound. The head tracking data was low-pass filtered in order to smooth the output data. Additionally, the transitions the convolved outputs were logrithmically crossfaded between the left and right channels to smooth the output signal which was sent to the auditor wearing the headtracker through stereo monitor headphones. Figure **??** shows the Max Patch.

IV. Results

The system was capable of binaurally approximating the auditory scene described in the Santala and Pulkki paper. The main limitations of this system was the convolution processing inherent in the MaxMSP programming environment. MaxMSP handles data through the signal vector size which determines the blocks in which it processes audio samples. This is typically set to be the DFT size (N) because of the prevalence of frequency domain processing in many of the native Max MSP objects. Initially with the 256 sample HRIRs forwarded to the FIR buffir filter blocks, the presence of around 8 virtual channels seemed to overload the computational processing involved using my CPU. After downsampling the HRIRs to 128, my computer was able to handle around 15 virtual channels.

The main drawbacks in the final system was there was a noticable output lag when the user quickly moved one's head to survey the auditory scene. Depending this velocity of this movement, a small lag of around 200-300 ms was noticeable enough to create a noticeable delay in the virtual sources which could be attributed jointly from the head tracking unit (discussed below) and the audio synthesis side of Max MSP. Additionally, there were occasional clips in the convolved output that were probably a function of the *buffir* objects becoming overloaded in trying to perform N channel simultaneous (linear) convolutions. Unfortunately, the buffir objects are unable to be modified given that they are native objects to the Max MSP environment.

Other ways to address this problem would be to employ time-domain based methods for the early portion of the impulse response and then use FFT-based approaches for the latter part (such as overlap-add) or circular convolution (overlap-save) using windowing which might speed up the processing considerably. However, because the HRIRs are comprised of a relatively small number of taps, its unclear if this would facilitate faster processing times and correct for the encountered time lags. While MaxMSP is advantageous because of its ease of use and availability across different platforms, one of the main limitations of using MaxMSP is the abstractions of its signal processing objects. It's also likely that this computational limit arises from the parallel, multi-channel nature of these HRIR convolutions being applied to the source signal. Since the time of this review, the HISSTOOLS Impulse Response Toolbox (Harker et al 2012) has come out which seems to have addressed the problems of real-time convolution and deconvolution in the current version of MaxMSP (v.7). Lastly, the head tracking unit that was used in this project was subject to a considerable amount of latency and noise. While data smoothing was attempted in MaxMSP, the precision and fidelity of the output azimuth angles would benefit greatly from better hardware-based filtering methods embedded on the unit itself.

V. Conclusions

This paper reviewed a preliminary method for using simulating point-source auditory scenes using an accelerometer based head tracking system and HRIR based convolutions delivered binaurally to stereo headphones. The system was shown to be able to account for head position in the context of 15 virtual speakers

References

[1] A. Harker, A. Tremblay, THE HISSTOOLS IMPULSE RESPONSE TOOLBOX: CONVOLUTION FOR THE MASSES in International Computer Music Conference, 2012

[2] L Ludovico, D.A. Mauro, and D. Pizzamiglio, HEAD IN SPACE: A HEAD-TRACKING BASED BINAURAL SPATIALIZATION SYSTEM , 2010
[3] J. Blauert,and W. Lindemann, Spatial mapping of intracranial auditory events for various degrees of interaural coherence , in Journal of Acoustical Society of America., vol. 79 (3), Mar. 1986
[4] K. Hiyama, S. Komiyama, and K. Hamasaki, The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field , Audio Engineering Society Convention., Oct. 5-8, 2002
[5] O. Santala, and V. Pulkki, Directional perception of distributed sound sources , Journal of Acoustical Society of America., vol. 129 (3), Mar. 2011
[6] V. Pulkki, Coloration of Amplitude-Panned Virtual Sources , Audio Engineering Society Convention., (110) May 2001
[7] A. Lindau, HRTF Datensatz fr die Horizontalebene Deutsche Telekom Laboratories, 2008