An Algorithm for Emulating Stereophonic Microphone Arrays

Jeffrey M. Clark, M.M.
Music Theory and Composition
School of Music, Ball State University

I. Overview

Audio localization in stereophonic recording focuses on balancing the pressure level and time difference cues that inform audio localization through interaural time differences (ITD) and interaural level differences (ILD). Within the stereophonic recording praxis, it is understood that the weighting of ITD and ILD cues contributes to varying desirable qualities within the resultant soundfield. Stereophonic microphone arrays are also chosen to account for the desired balance of direct sound and reflected sound and to accommodate the recording angle of the perceptual "soundstage" and translate it to audio reproduction systems with minimal angular distortion.

This is, however, at odds with typical stereophonic localization practice within the application of audio signal processors; which tend to prioritize localization solely through ILD cues. In certain contexts this can lead to a situation where recorded signals that are being mixed together during post-production can be localized using different perceptual locative cues, yet with the intent of producing a coherent, and qualitatively consistent perceptual soundstage.

By modelling the time and level relationships between a sound source and the microphones in a stereophonic recording array, a monophonically recorded sound can be localized within the perceptual soundstage using the time and level cues appropriate to the modeled array. This paper will outline an algorithm and discuss and implementation for abstracting these relationships in to a CPU-efficient model, and exposing the appropriate parameters to the user.

A. A Review of Stereo Recording Localization

Within the context of localization, it is understood that a relative dominance of time-domain cues creates a greater perceived sense of stereophonic width/envelopment, whereas a greater relative preponderance of level-based cues increases the sense of locative precision. Microphones, as pressure transducers, encode local changes in atmospheric pressure into changes in electrical pressure, and can be practically thought of as sampled points in space. Additionally, microphones have varying types of directivity – represented by a polar plot – that represent how efficiently they transduce sound based on the sound-wave angle of arrival relative to the microphone's orientation. For directional microphones with cardioid-style polar patterns, the amount of attenuation tends to increase as the sound-wave's direction of arrival moves further away from the oriented "front" of the microphone[1].

By positioning two directional microphones in the same location and facing in two different directions, any sound-wave captured by them will be increasingly less attenuated as it approaches the front of one microphone and more attenuated as it approaches the other. If the sound wave approaches from an angle that equiangular to both microphones, then it will be equally attenuated. This equal attenuation, during reproduction, will have the perceptual effect of placing the sound at the center of the virtual sound-stage.

Similarly, by positioning two nondirectional microphones in to locations within the space, the distance between them will create differences in the time of arrival based on the speed of propagation of the sound-wave (the speed of sound). As the direction of arrival shifts away from being perpendicular to an imaginary line drawn from one microphone to the other, there will be an increasing time delay between when the sound is captured in the closer microphone and when it is captured in the further microphone. As with the level-based cues, when the time of arrival is equal (i.e. there is no delay in one microphone) then the perceptual effect during reproduction is one of the sound being centered in the perceptual soundstage. As the delay increases in one microphone, the sound will appear to come from the direction in which it arrives first[2].

These level and time-based principles can be freely combined, creating a spectrum of options for recording engineers to choose from. More advanced methodologies will add in a center microphone (such as the popular "Decca-Tree" configuration), and/or will also add flanking/outrigger microphones as well. The combined effect of these various microphone set-ups yields a complex interplay of time and level cues.

[1] There is some complexity with this as the directivity pattern value increases past a pure-cardioid, with an inverse-phase area beginning to present at the rear of the microphone and increasing until the directivity approaches a bidirectional pattern

[2] This phenomenon is known as the "precedence effect"

B. Signal-Processing Panning Methods

Current methods in stereophonic panning within digital audio workstations (DAW) typically focus on the manipulation of level differences. These level differences tend to follow either a sine-cosine (see 1) or a linear curve (see 2).

For $\{p \mid 0 \leq p \leq 1\}$ to represent the range of the user input value for stereophonic panning, then the normal panning functions can be found as:

$$\begin{bmatrix} L \\ R \end{bmatrix} = x * \begin{bmatrix} \cos \frac{p\pi}{2} \\ \sin \frac{p\pi}{2} \end{bmatrix} \tag{1}$$

$$\begin{bmatrix} L \\ R \end{bmatrix} = x * \begin{bmatrix} p \\ 1-p \end{bmatrix} \tag{2}$$

Interestingly, there is not as accepted standard functionality for DAWs to perform panning based on time-domain cues. Multiple reasons can be found for this, though the two that arguably stand out are: lack of monophonic compatibility due to phase distortions, and tradition from analogue mixing consoles which were unable to provide the delay lines necessary. Additionally, the usage of computer memory for the circular-buffer delay-lines needed may have also been a consideration against memory availability in early computers.

The use of time-domain cue is either implemented through a stereo "widener[3]" or by manual implementation through simple or purpose-built delay audio plug-ins in a technique known as the "Haas Trick[4]."

C. Purpose

This paper will describe an algorithm for modeling the interplay of the microphones within a stereophonic recording array; and discuss an audio plug-in that implements this algorithm. This algorithm and the accompanying audio processor were developed as an entry to the AES/MATLAB Plugin Competition in 2020. As such, the processor developed was designed to work within the limits of the competition – most specifically the two-channel input/output requirement, and the goal to solve a specific use-case scenario. Theoretically, this algorithm could be applied to any arbitrary microphone set-up and source placement within a virtual space.

[3]The implementation of stereo widening techniques varies from processor to processor, not all will use the same methods. Any given widener may not actually use time-based cues, relying instead solely on frequency-domain adjustments or dynamics range processing.

[4]after Helmut Haas, who studied the psychoacoustical implications of the precedence effect in his Ph.D. thesis.

The immediate use-cases intended to be addressed with this specific implementation are: 1. the placement of monophonically recorded sources into a stereophonic microphone array, and 2. the addition of time-domain cues through a UI that is intuitive and effective, and yields natural and predictable results.

1) Blending monophonic encoded sources: When recording acoustic ensembles, recording engineers will frequently employ both a main stereo array and area/spot microphones. These spot microphones are used to accentuate various parts of the ensemble to difference logistical and aesthetic ends. Drawing from previous sections, stereo recording praxis and panning implementation in DAWs do not agree on how to encode stereophonic localization cues. This can quickly lead to a situation where the sound of an instrument recorded in an ensemble with both a stereo array and spot microphone will have a mixture of time/level cues in the main array signal, but only be localized with level cues in the panned signal from the spot microphone. Due to the differences in the perceptual effect of time and level cues, this can lead to inconsistencies in the quality of the soundfield as the dominance of the source of the instrument in the audio mix changes between the main array and the spot microphones; this may also cause inconsistency in instruments that are around the spot microphone. The use of this processor will encode the monophonic sound with time and level cues that closely match that of the main array, meaning that its localization and sense of envelopment will not change based on the dominance of the source.

Following this, another context for this application of the processor is in blending a monophonic recorded source into a recording that was taken of an ensemble using a stereophonic microphone array. "Distributed recording" is not uncommon – especially in music for media. It is not uncommon for sections of an ensemble to be recorded independently and then put together during post-production. This processor would allow for independently recorded instruments to be better blended into a larger ensemble by encoding them with the same localization cues, creating a greater qualitative unity in the resultant soundfield.

This also has implications for sample-based music production and the common practice of laying samples of the same instrument family that come from different sample libraries. Commercial sample libraries each have their own recording and staging procedures, and different instruments and sections can be localized using different combinations of level and time-domain cues. The use of this processor on the spot microphone or monophonic summed options will allow for the sounds to be spatialized together in a more consistent manner.

Finally, this processor can also be used in conjunction with standard post-processing artificial reverberation techniques to manufacture a sense of unity from completely independently recorded instruments by encoding them onto a virtual soundstage using natural level and time-domain cues.

2) Time-domain cues in panning: A secondary use for this processor is in presenting an intuitive method of adding time-domain cues into panning. Rather than relying on an arbitrary implementation of the Haas trick, a user of this processor can – through the familiar metaphor of a stereophonic microphone pair – creating a consistent, repeatable, and intuitive panning plan for components of their audio mixdown that includes both time and level perceptual components.

II. Definitions

A. The Sound Source

1) The propogation of sound from real sound sources: The first immidiate barrier to modeling the propogation of sound from a sound source is the nature of any given sound source. Few acoustic resonators approach being perfect point sources with equal spherical radation of sound accross the entire audible frequency spectrum.

III. Positional Relationships

IV. Virtual Microphone Processing as a Complex Number

V. Δt Compensation

VI. Approaches to the Abstraction of the Damping of Sound Pressure

References

[1] R. King, *Recording Orchestra and Other Classical Music Ensembles*. New York City, NY: Routledge, 2017.

[2] M. Williams, "The stereophonic zoom," tech. rep., Microphone Data, 2010.

[3] M. Williams and G. L. Du, "Multichannel microphone array design," tech. rep., Microphone Data, 2010.

[4] D. Arteaga, "Introduction to ambisonics," tech. rep., Dolby Laboratories, Inc., 06 2015.

[5] J. Meyer, *Acoustics and the Performance of Music*. Modern Acoustics and Signal Processing, Springer, 5th ed., 2009.