2020 Sound

by

Finlay Braithwaite

A thesis exhibition presented to OCAD University
in partial fulfillment of the requirements for the
degree of
Master of Design
in
Digital Futures

Toronto Media Arts Centre, April 12 - 14 2019

Toronto, Ontario, Canada, April, 2019

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I authorize OCAD University to lend this thesis to other institutions or individuals for the purpose of scholarly research. I understand that my thesis may be made electronically available to the public. I further authorize OCAD University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Signature _____

Abstract

Finlay Braithwaite, OCAD University, 2020 Sound, Master of Design (MDes), Digital Futures, 2019

2020 Sound is a positional-tracking microphone and DSP time-of-arrival alignment system. In production, an ultrasonic beacon emits a temporal positional reference that is captured by standard audio recording devices. This reference gives purchase to align multiple microphone perspectives of a source, correcting for their initial offset as well their movement throughout the recording. In capturing a sound source with multiple microphones, misaligned and drifting time-of-arrival of the source at each microphone greatly impacts the cohesion, focus, and impact of their summation in the mixing process. The common boom and lavaliere microphone scenario implemented in film and television production suffers from this misalignment and, as a result, time-intensive and inaccurate manual editorial processes are employed to align microphones before their summation. This system could remedy a fundamental issue encountered in audio production with the ultimate aim of improving the clarity and quality of productions that make use of 2020 Sound.

Acknowledgements

I couldn't have done this without the love, support, patience and understanding of my family, my wife Justyna and my son Ernest. My parents Deborah and Colin and my brother James also provided essential help that made this possible. Thanks to my friends and colleagues at Victory Social Club and Ryerson University. Finally, thanks to my advisors Nick Puckett and Adam Tindale for their support and guidance.

Table of Contents

Table of Figures

1 Introduction

This research through design project explores possible solutions for the alignment of time-of-arrival for a sound source arriving at multiple microphones. Scaffolding this research is an exploration of the current narratives of time-of-arrival in cinema audio production and post production contexts. The impact of the 2020 Sound design solution is superimposed on these existing narratives, highlighting the intrinsic and extrinsic potentials of TOA alignment. These narratives are a synthesized amalgam of my own professional experience in the field and the input of experts from audio production, post production, and engineering technician backgrounds.

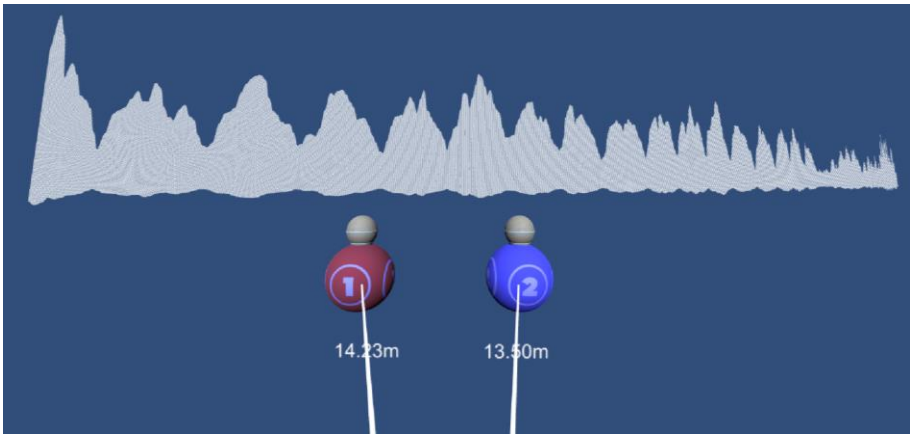

Figure 1 - Constructive and Destructive Interference of Two Summed Microphones Capturing a single source.

The problem of time-of-arrival of multiple microphones negatively impacts the work I do as an audio professional. Every time I bring up two microphones in a mix, I lose as much as I gain when combining their images. They work against one another, softening and blurring the potential that each microphone has on its own. This is a constant challenge, forcing me to use one microphone instead of multiple; only reaping the benefits of one perspective while sacrificing another.

While the issues of time-of-arrival are fixed and captured in production, they manifest in full during the post production process, limiting the mixer's ability to leverage the combined potential of all microphones captured in production, often tying their hands to implement only one microphone in the final mix. Through this work's exploration of the narratives of postproduction, the rationale for aligning time-of-arrival is revealed while highlighting the potential benefits of a solution.

The novelty of these explorations lies not in the aim of aligning time-of-arrival but rather in its approach. This work acknowledges existing techniques with parallel objectives as well their narrative in current production and post production workflows, evaluating the potential and limitations of these existing approaches.

This research explores the analogous problem of wow and flutter in mechanical analogue mediums as well as the digital latency inherent in complex audio systems. In the case of wow and flutter, this examination of comparable temporal distortions connects to a wealth of possible solutions that could be applied to the microphone time-of-arrival alignment problem.

Contemporary designs that add focus and clarity to audio productions illustrate the context in which this work resides. While some designs examined do correct temporal distortions including time-of-arrival across multiple microphones, the common element of all examined designs is their interventions in the minutiae of perceived audio quality. While situating the work in a larger context, this exploration punctuates the extrinsic purpose of this work, to play a role in improving the potential for clarity, focus, and precision in creating the next generation of soundtracks.

This work stems from my experience in professional sound production and post production. Complimenting the professional experience is my passion for teaching audio production to the next generation of sound professionals. These narratives, in combination with a survey of experiences of the community of sound professionals, serve to define the problem while providing insight into its potential

solutions. An iterative approach to research through design fixes these narratives into the functional prototype produced.

In developing the functional elements of the design, their potential in the service of capturing, documenting, and aligning the position and temporality of sources and microphones were explored. This work looks into the technical building of the blocks for a solution that spans both production and post production contexts. Not all elements detailed are used directly in the finished designs as some components inspired the ultimate solution and others fulfilled the role of catalyst in exploring possible solutions.

Beyond the technical underpinnings of the proposed solution, the user interaction and experience of this solution were explored from production and post production perspectives. How users connect to the potential of TOA alignment was examined alongside potential barriers and challenges that the design attempts to mitigate.

Ultimately, this work culminates in a functional prototype of an ultrasonic clocking and digital signal processing time-of-arrival alignment system. The result is a virtual proof-of-concept prototype that lays the foundation for future iterations of the design. This acts as a foundation to further develop tools to align time-of-arrival with the ultimate goal of improving the perceived quality and impact of soundtracks.

2 Properties of Sound

As this investigation explores the minutiae of sound in space, it would be remiss to not include a primer on the properties of sound that support and direct these investigations. From this framework of the fundamental properties of sound, the inherent issues of multiple microphone perspectives of a single source are revealed. This section then explores the challenges and potentials of multiple microphone techniques in production and postproduction. With the problem identified, the analogous problem of wow and flutter in analog recordings is examined for its parallels and possible insight into the time-of-arrival alignment problem at the core of this work.

2.1 Sound Fundamentals

Sound is objects vibrating in a medium. A vibrating object cyclically pushes and pulls on the surrounding medium creating alternating states of compression and rarefaction. These compressions and rarefactions radiate outwards from the object in all directions. The magnitude of the compressions and rarefactions defines the amplitude of a sound. Their rate defines the frequency of the sound.

The speed of sound radiating away from the source is highly dependent on the medium itself. In our atmosphere, the speed of sound is generalized as 343 m/s[1]. This value assumes a temperature of 20 degrees Celsius with the speed of sound increasing with temperature. The composition of the medium also plays a large role in the speed of transmission as, for example, sound travels 4.3 times faster in water than in air[2].

[1] "Speed of Sound." Wikipedia. March 08, 2019. Accessed March 18, 2019. https://en.wikipedia.org/wiki/Speed_of_sound.
[2] "Speed of Sound." Wikipedia. March 08, 2019. Accessed March 18, 2019. https://en.wikipedia.org/wiki/Speed_of_sound.

2.2 Frequency and Wavelength

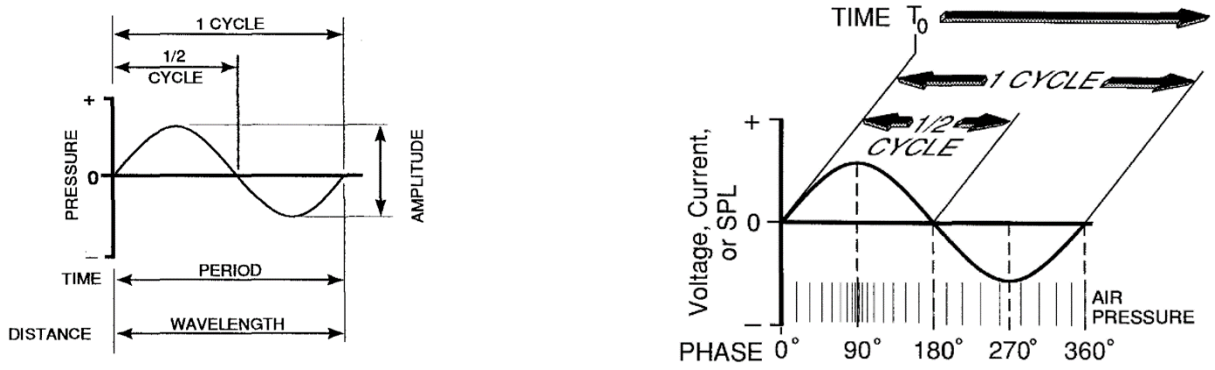

Figure 2- Amplitude, Phase, and Wavelength of Sound[3]

The cycles of compression and rarefaction can be measured both in terms of time and space. The length of time it takes for a full cycle of compression and rarefaction is its frequency, measured in cycles per second or Hertz(Hz). The physical length of the cycle is documented as wavelength, measured in metres. As frequency increases, wavelength decreases and vice versa. This connection of time and space in the physical properties of sound are a fundamental component of the investigations of this project. At any given point in time, a sound's current state of compression and rarefaction can be described as its phase, measured in degrees. The beginning of a cycle being 0 degrees and the completion of cycle documented as 360 degrees.

2.3 Time of Arrival

As sound radiates from a source at the speed of sound, it arrives at objects in the acoustic space in order relative to the distance of each object to the source. A close perspective would have an earlier time-of-arrival relative to the later time-of-arrival of a more distant perspective.

[3] Gary Davis and Ralph Jones, *The Sound Reinforcement Handbook*, 2. ed., 2. printing (Milwaukee, Wis: Hal Leonard, 1990).

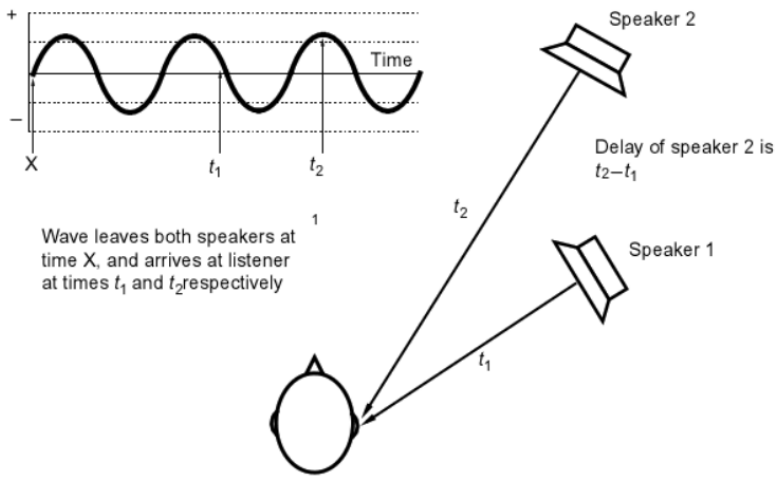

Figure 3 – Time-of-arrival of two sources with common signal[4]

If a source is pushing against the atmosphere - creating a compression, the time-of-arrival of that compression can be calculated by factoring the speed of sound against the distance of source to destination. If a source is 343 metres away from a destination, the sound arrives 1 second later.

Put simply, for the purpose of this thesis, time-of-arrival is the amount of time it takes for a sound to arrive at a destination from the source of sound itself. In this context, the destinations are microphones that transduce the acoustic energy into an electrical current, allowing for its recording.

2.4 Reflections

As sound transverses an acoustic medium it is absorbed and reflected by the objects in its path. From the point of reflection, these reflection in turn then radiate within the acoustic space. Therefore, perception of a sound source is a combination of the direct sound from the source as well as reflections of the source. As the path of reflected sound is not a direct line between source and ultimate point of perception, the time-of-flight is subsequently longer. At the point of perception, the direct sound has an

[4] Francis Rumsey and Tim McCormick, *Sound and Recording: Applications and Theory* (Oxford, UNITED KINGDOM: Taylor & Francis Group, 2014), http://ebookcentral.proquest.com/lib/ryerson/detail.action?docID=1638630.

earlier time-of-arrival compared to that of the reflection. The rich confusion and softening of a sound source through reverberation is an analogous phenomenon to the summation of multiple microphones with varying time-of-arrival.

2.5 Summing multiple perspectives

If summing multiple perspectives of a single source, the output of multiple microphones for example, the combination is at times additive or subtractive depending on the state of compression or rarefaction of each mic at any point in time. One microphone may be transducing a compression whereas another transduces a state of rarefaction. In this transduction, a compression translates to a positive voltage whereas a rarefaction results in a negative voltage.

Whether a summation is additive or subtractive depends on the frequency or wavelength of the source and the distance between the microphones used in the summation. If the distance between perspectives is an exact multiple of the source's wavelength, the summation is completely additive. The spaced pair of microphones in this case would be described as being in phase. A spacing that lies precisely halfway between one of these multiples would be totally subtractive, resulting in a summation result of zero or silence. Perspectives with this spacing would be considered out of phase. Points between in between these extremes lie on a gradient of additive and subtractive summation results.

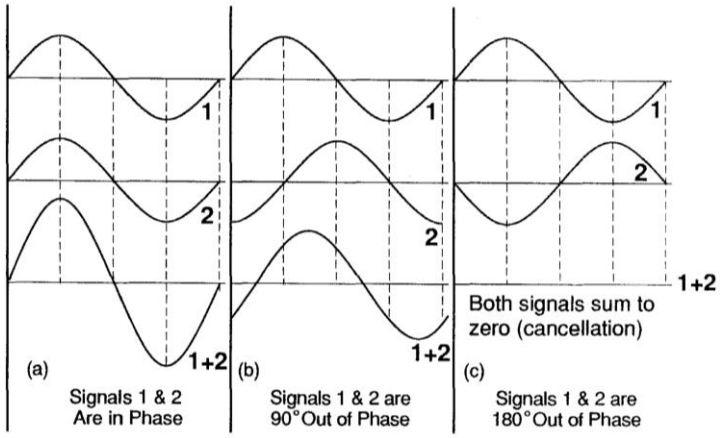

Figure 4- Summation of sine wave in varying phases[5]

2.6 Comb filtering

This talk of phase is overly simplistic in that it supposes sound sources vibrating at a single simple frequency, a sine wave. Practically speaking, the vibration of objects in space and the sound they generate is more complex. While a sound may have a fundamental tone, more often than not there are complex overtones and harmonics the prevent the sound from being described or documented as single frequency. Rather, a complex sound generates multiple frequencies or an entire spectra of frequencies.

If a sound source can be conceptualized to contain multiple frequencies, the additive and subtractive nature of the summation of multiple perspectives is much more complex. Across the entire frequency spectrum, the distance between perspectives relative to source has a varying effect on whether the summation of a frequency is additive or subtractive. Alternating bands of the spectrum emerge as additive or subtractive, creating a comb-filtering effect.

[5] Davis and Jones, *The Sound Reinforcement Handbook*.

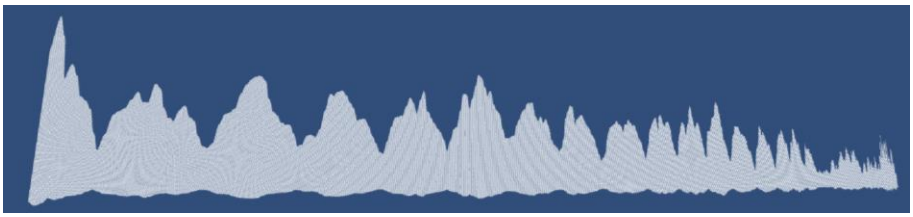

Figure 5 – Comb-filtering resultant of two microphones at varying distances from source

2.7 Spatial Audio

As ears are separated by the dimensions of the head, one ear leads the other in time-of-arrival of a source, depending on which ear is closer to the source. This is one element that defines the head-related transfer function (HRTF) that allows humans to positionally locate a sound source in space. Another key factor of HRTF is the impact that the mass of the head has as it shades the ears from the source in terms of amplitude and frequency response, the emphasis and attenuation of frequencies depending on the head's position relative to the ear.

The core of this investigation probes the relationship of multiple microphone perspectives of singular sound sources. While there is a connection between this work and spatial audio, it is a tangential connection through aural perspective in real and virtual spaces. Spatial audio, in its attempts to create hyperreal reproductions for its audience attempts to map perspective of sound fields and sources to our two microphones, our ears. These two acoustic transducers are a spaced pair, but everyone's spacing is unique. The holy grail of spatial audio research is translating the head-related transfer functions of one spatial experience to another. Suffice to say, the position of microphones, so to speak, is essential to the domain of spatial audio. It is an important distinction that this work is primarily concerned with relative distance between sources and microphones rather than absolute coordinates of a source in three-dimensional space.

Nils Peters' work explores systems of virtualizing acoustic space so that translations can be made to facilitate accurate HRTF experiences for the audience[6]. In this virtualized space, virtual microphones can be synthesized to capture the positional and directional perspective at particular coordinates in the virtual space. Peters' investigations holds potentials to this thesis work as it endeavors to virtualize space in order to investigate possible solutions to align time-of-arrival of multiple microphones. In audio, time is distance - tied to one another by the speed of sound. In altering the timing of multiple microphone perspectives, it's changing their position in a virtual sense, creating a virtual world where alignment of microphones is just one possibility.

[6] Nils Peters, "Sweet [Re]Production: Developing Sound Spatialization Tools for Musical Applications with Emphasis on Sweet Spot and off-Center Perception," n.d., 305.

3 Time of Arrival Narratives

Time-of-arrival is a fundamental component of the relationship between sound sources and perspectives. As such, it play a large role in the transduction, capture, and ultimate production of sound media. Its impact on these processes is unavoidable and, as a result, becomes a fundamental consideration of producers of sound media.

This work explores the narratives of time-of-arrival from the perspective of audio production and post production perspectives. Through the collection and synthesis of these narratives, the rationale for intervention and ultimate alignment of time-of-arrival emerges and becomes apparent. These narratives stem from my professional experience and are scaffolded through interviews conducted with production and post production audio professionals in film and television.

3.1 Production

At its core, the production scenario is the transduction of live audio sources in an acoustic space. While time-of-arrival is an important consideration in most audio production scenarios, the paradigm of film production sound production is the focus of this work. Regardless, there are parallels of this narrative to radio, television, music, and theatrical scenarios.

In the film production sound scenario, the primary objective is to capture actors' dialogue with as much presence, clarity, and focus as possible.

Multiple microphones are used to capture actors as they perform. A typical scenario could engage boom, body, and plant microphones to capture a singular source.

The boom microphone is typically suspended above the subject, placed as close to the actor as the camera's framing allows. The boom is a long arm that extends the reach of a boom operator, allowing the placement of the microphone above the scene. The boom itself can range from a hand held pole to

a long fixed stationary pole with mechanical controls for movement. To isolate the subject from other sources in the scenario such as extraneous sounds, other subjects, or the acoustic of the space, the microphones used in boom operation are highly-directional. The boom operator's challenge is to maximize proximity and point their microphone towards the subject, all the while avoiding appearing in the camera's frame. This relationship between boom and camera creates a sympathetic match in perspective between the visual and aural capture. With a close-up visual composition, the microphone can be very close to the subject. With a wider shot, the boom must be further away.

The body microphone is a small microphone known as a lavaliere microphone. While a lavaliere can be affixed in a number of ways, the film production scenario requires it to be hidden on the body of the actor. As with the boom, the appearance of the microphone in-frame would break the diegetic narrative of the film by exposing the apparatus of the production. In this production context, the lavaliere microphone is omni-directional. The microphone, therefore, can only be focused towards or away from a source through its proximity. The body microphone is taped to the actor or integrated into their wardrobe. As such, the microphone's perspective of the source is fixed resulting in a constant perspective that does not sympathetically parallel the visual shot composition as the boom microphone does.

The body microphone, partially due to its size, is often a lesser quality microphone in comparison to a boom or plant microphone. To allow for movement in a scene and to prevent the cabling of the microphone from being seen by the camera, a wireless system connects the microphone to a recorder. This wireless transmission can only have a detrimental impact on quality when compared to a cabled boom or plant microphone. The positioning of the microphone, often on the actor's sternum underneath their clothing, is reflected in its perspective, a somewhat muffled perspective of a voice that is shaded by the chin. It is for these reasons of perspective and sonic fidelity that result in body microphones' consideration as secondary or backup microphones to boom or plant microphones.

A plant microphone is similar to a boom microphone in many ways but is physically hidden in the scene rather than being suspended from above. These microphones can be planted in props or furniture that are proximate to an actor. In a car scenario, the microphone can be hidden in the car's visor. In an office scene, the microphone could be hidden on the actors desk. Not all shots present opportunities for hiding microphones, but they can be a useful complimentary microphone to the boom and lavaliere. As fixed microphones, their perspective is dictated by their position and directionality and do not respond to the changing visual frame of the camera nor the movement of the actor.

It is through this implementation of multiple microphones of a single source that issues of time-of-arrival arise.

If two microphones capture they same source, the time-of-arrival of that source will be misaligned between the microphones. The closer microphone transduces the compressions and rarefactions emanating from the sound source first, with the transduction of the same compressions and rarefactions happening later in more distant microphones. It is this scenario that is the primary focus of this work: the misalignment of multiple microphones due to the relative differences in distance between a singular source and multiple microphones.

This misalignment is apparent in the mix created by the production sound mix operator. As they combine the misaligned images of multiple microphones into a production mix, the mix loses focus, clarity, and precision.

3.1.1 Two Sources, one perspective

There are, however, other time-of-arrival scenarios that impact the production process. While this work focusses on the scenario of a single source perceived at multiple perspectives, time-of-arrival has the same effect in the inverse scenario of multiple matched sources arriving at a single perspective. This inverse problem has parallel narratives in production and post production.

As an example, take the aural experience of a boom operator. They extend a microphone with a boom, placing the microphone closer to the source relative to their ears. The boom operator perceives a confused image that is the blend of the earlier arrival of the boom microphone and the later arrival of the same source at their ears.

3.2 Post Production

Post production in this context refers to the audio editorial and mix processes that create the final soundtrack for a film. These processes occur after the narrative video editorial process where audio and visual material captured in production is selected, edited, and sequenced to create the fixed linear form referred to as locked picture. As the focus of this work is the alignment of multiple microphones, this section focuses on the dialogue editorial and mix processes specifically. Generally, dialogue is the primary consideration and focus of capture in production. Other elements such as sound effects and music are added entirely in the post production process.

Time-of-arrival is not a primary consideration of the dialogue editorial team unless they endeavour to correct or align using manual methods detailed in subsequent sections.

The focus of the dialogue editorial process is to finesse and augment the dialogue cut from the narrative video editorial process. In evaluating the recorded dialogue, the dialogue editorial team smooths the edits of the video editor and, if it cannot be finessed, replaces the selected line with another take or performance from production. If no fix can be found, the actor is reengaged to reperform the line in a studio as a replacement to the original dialogue in a process known as ADR. Extraneous noise and sounds contaminating a recording or an issue of talent performance are common obstacles that require a line to be rerecorded and replaced.

The dialogue editorial finesses and delivers edits of all microphone types recorded in production, leaving the decision of which microphone to implement to the downstream mix process. Even in rerecording

performances through ADR, multiple microphones are recorded to mimic the perspectives recorded in production, thus allowing for a more seamless match between production and replaced dialogue recordings.

It is in the mix process that time-of-arrival becomes apparent as an issue. As all microphones are finessed and passed from the editorial to mix stage, the mixer has options as to which microphones contribute to the final soundtrack.

A mixer combining multiple microphones is presented with the challenge of combining multiple perspectives with varying time-of-arrival. As outlined earlier, the summation of multiple microphones on a single source is not wholly additive, nor wholly subtractive. As the microphones blend with one another, common characteristics of the two perspectives are emphasized and differences are deemphasized. The time-of-arrival differences between the two microphones combine to create a blurry, smeared image with less focus and fidelity than that of each discrete perspective.

With the problematic reality of mixing misaligned microphone positions, the most precise option for mix clarity and focus is to use only one microphone. This removes any lack of clarity introduced by mixing multiple microphones. In this case, the boom microphone would be the ideal choice with its sympathetic match between the visual and aural perspectives. If the boom is not a feasible choice, say for reasons of poor or inconsistent boom technique, other mics may be selected as the sole contributor to the mix. However, with either body or plant microphones, aural perspective would have be matched to the visual through labour-intensive processing and automation. Otherwise, the sound doesn't respond to the visual perspective and feels disconnected as a result.

3.2.1 Time of Arrival Alignment

Aligning time-of-arrival restores the potential of contributing multiple microphones without suffering a loss in fidelity in doing so. With aligned microphone perspectives, a mixer is able to blend perspectives

as desired. For example, the shifting perspective of the boom could be grounded and reinforced with the consistency a lavaliere. Two microphones that before worked against one another would become more complimentary. The sonic potential of mixes are expanded as they allow for the contribution of elements from multiple perspectives rather than that of a single microphone.

3.2.1.1 Current Methods

These benefits are not the stuff of theoretical abstraction. A method of alignment for multiple microphones exists and is implemented in productions that have the expertise and capacity to implement it. However, this method is manual and, as a result, is time and labour intensive. Resultingly, it is not a commonplace practice.

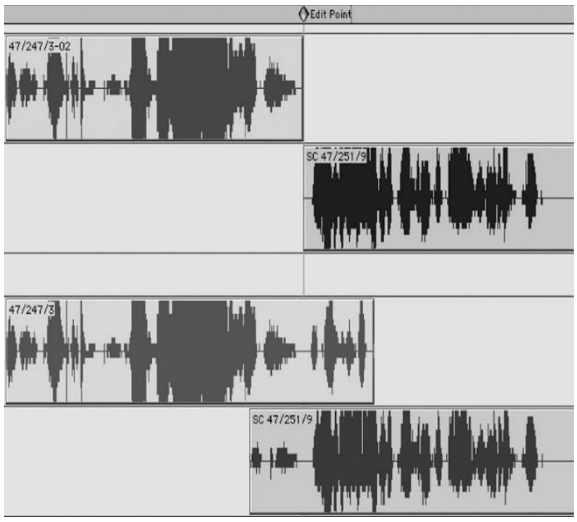

Figure 6- Dialogue Editorial Alignment[7]

[7] John Purcell, *Dialogue Editing for Motion Pictures : A Guide to the Invisible Art* (Routledge, 2013), https://doi.org/10.4324/9780203784570.

To align perspectives of multiple microphones, the recordings are manually cut and aligned in editorial. By eye, an editor looks at the cycles of compression and rarefaction and advances or delays them in order to align them. This alignment is static and does not allow for the dynamic alignment of moving microphones or sources. The resolution of the alignment increases with each additional cut and alignment but can never fully align against continuous movement.

3.3 Analogous Problems

3.3.1 Issues in Music

In music production, it is common to implement multiple microphones to capture a single source or a combination of sources. Take for example, contemporary practices of capturing a drum kit.

3.3.1.1 Capturing Drums

While a single microphone captures the drum kit overall, it does not allow for the discrete rebalancing of drums post-production in the mix. A kick drum cannot be raised in amplitude relative to a tom drum, for example. As such, individual microphones are used to capture each individual drum as discretely as possible. While it is possible to create a drum mix using only these discrete perspectives, aesthetically this mix lacks the cohesion of the kit as a whole. A compromise is struck when overhead microphones are implemented to capture the drum kit as a whole while individual spot mics are used to capture individual drums. The perspective of this approach can treated either way in the mix as the mixer can select which microphone type, spot, or overhead form the basis of the mix and which type acts as reinforcement.

As the spot microphones are relatively closer to each individual drum, their time-of-arrival is earlier than that of the overhead microphones. Mixing the two perspectives has the overhead microphones behind the spot microphones, softening the transient impact of the percussive drum strikes. Striking a drum is a

highly percussive sound with a strong transient or sudden change in sound pressure level. It's worth noting that this softening of the drums can be an aesthetic choice in its own right.

A common strategy to align the time-of-arrival between overhead and spot microphones is to delay the playback of the spot microphone recordings in the mix. In delaying the spot microphones, they are effectively moved in space, delaying their time-of-arrival. Through this method, the time-of-arrival of spot microphones can be aligned to that of the overheads. This approach is only effective when delaying the spot microphones in that, as with their amplitude, individual drums cannot be retimed in the overhead microphone image. By aligning time-of-arrival in the mix, a punchier drum sound is created through a single non-diffused transient.

With or without this alignment, it a common approach to gate drum microphones. The process of gating effectively turns a microphone off if its target drum is not in use at a given time. By turning microphones off when they are not in direct use, they do not contribute off-microphone bleed to the mix. For example if someone is playing the snare and not the tom, the tom microphone turns off, cancelling the downstream time-of-arrival perspective of the snare on the tom microphone. Through this technique, the mix is made more coherent by having the fewest possible microphones contribute to the mix.

Another approach is to mitigate the issue of time-of-arrival in production. The close microphone approach relies on several microphones. Techniques that use fewer microphones naturally have fewer intersections and conflicts of time-of-arrival. A common technique is to use three microphones that are equidistant to the centre of the drum kit, the snare. This results in recordings that are fundamentally more aligned than the overhead and close microphone approach.

While it is not the explicit goal of this work to align audio in a musical context, the proposed design would accommodate this scenario, resulting in mixes that retain more of the transient punch of their acoustic sources.

3.3.2 Wow and Flutter in Analog Recordings

The theoretical solution and prototype for the analogous problem of wow and flutter on vintage recordings holds potential in the investigation into time-of-arrival alignment of multiple microphones.

Many old recordings are warped; they speed up and slow down, pitching up and down as a result. Recording audio is a process of capturing a dynamic electronic signal that documents a musician, speaker, or whatever source occupies the content dimension of the recorder's input. This material is recorded against time, correlated against the recording medium as a transport rate. For example, analogue tape would specify its transport rate in inches per second, connecting time to the medium. Each recording and playback device relies on a clock to drive this rate. For example, your turntable may have a quartz clock circuit that drives the physical turning of a platter.

Ideally, a device's clock would be perfect and material recorded and subsequently played-back on that machine would perfectly reflect the temporality of the recording scenario. However, these clocks and the devices they drive are not perfect. All audio devices deviate from their specified transport rate to varying degrees. Older mechanical analogue devices such as tape and vinyl suffer from wow and flutter, the relatively low and high rate audible deviations of transport rate respectively. Contemporary devices suffer from jitter, variations of a higher-order that while not directly audible, compromise the quality of an audio recording.

If a recording device suffered from noticeable wow and flutter in recording, it's baked-in to the recording, manifesting as the inverse deviation on playback. For example, if a recorder transport rate slowed during recording, there would be an audible increase in rate and pitch during that segment. In vinyl recordings, the clock of the original recording could be distorted with misaligned centre hole.

This temporal distortion has parallels to the doppler effect perceived as a microphone moves closer relative to a source or vice versa. In the case of analog recordings, the modulation is the result of a bad

clock that drives a mechanical transport mechanism whereas the distortions investigated in this work are the result of changes in distance and the propagation of sound waves at the speed of sound.

In one phase of their work to correct wow and flutter, Wallaszkovits et al leverage a hidden constant buried in the analogue tape recordings containing wow and flutter[8]. The recordings contain a high-frequency bias tone that resides in the inaudible ultrasonic frequency domain. This tone, created by an oscillator (clock) at the genesis of the recording, originally had a consistent pitch. Recorded to tape, this tone is no longer constant but suffers from the same wow and flutter of the audible recorded material. Wallaskovits et al implement a process to restore this fluctuating pitch to a consistent pitch and, in the process, remove the wow and flutter from the recording as a whole.

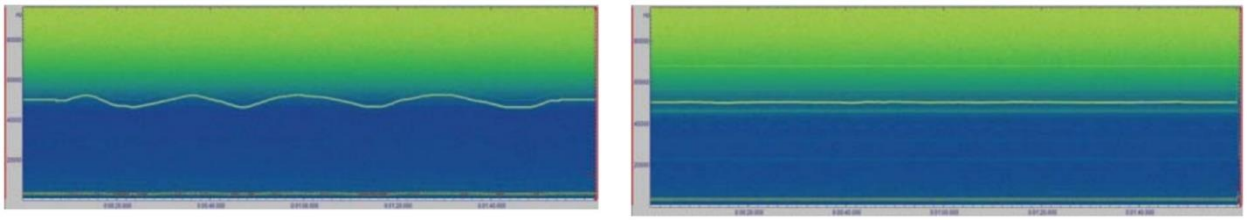

Figure 7 – Wow and Flutter Bias Tone: uncorrected (left) and corrected (right)[9]

If a hidden clock can be recorded alongside normal content, hidden from the audience, then this clock could be used to document TOA and facilitate its correction. In Wallaskovitz et al work, the figurative high-frequency bias 'bread crumbs' that allowed them to return recordings to a flat temporality were an unintentional consequence of the original recording technology implemented. Intentionally creating and documenting a similar inaudible frequency provides purchase to align multiple moving microphone perspectives to a source.

[8] Nadja Wallaszkovits, Tobias Hetzer, and Heinrich Pichler, "Drift and Wow Correction of Analogue Magnetic Tape Recordings in the Analogue Domain Using HF-Bias Signals," in *Audio Engineering Society Convention 136*, 2014, http://www.aes.org/e-lib/browse.cfm?elib=17189.
[9] Wallaszkovits, Hetzer, and Pichler.

Walloskovitz et al are not the first or only to investigate the flattening of wow and flutter. Andrzej Ciarkowski and Andrzej Czyzewski have an extensive body of work in the pursuit of correcting warped analog recordings. In their work, low frequency hum present in many analog recordings presents a temporal guide for restoration[10]. While on opposite ends of the audible spectrum, the principle is similar; using an artifact from the recording process as way to correct pitch deviations.

3.3.3 Audio System Latencies

Complex audio systems suffer from the impact of computational latency. This latency is the result of digital processes that digitize, buffer, and process audio. Each process takes time and therefore adds an offset or delay to the signals passing through them. Contemporary digital audio systems compensate for their internal delay, ensuring that, while there is an overall delay, all signals are delayed equally regardless of the delay required for each signal. In Avid's Pro Tools, for example, all signals are delayed to match the longest or most latent processing path in a process known as delay compensation. In playback, this delay is a non-issue as the recordings that feed playback for processing are advanced to negate the delay.

It is in larger interconnected digital audio systems with multiple independent devices that this problem emerges on a larger scale. Each device has its own latency associated with it. As the systems are unique and often proprietary, there is little hope of delay compensation being implemented in an effective way.

This misalignment is not necessarily static either. As sample rate convertors allows unclocked digital devices to be connected to one another, they are allowing for the interconnection of devices that are running at fundamentally different rates. This creates dynamic temporal distortions between devices connected in this fashion.

[10] Andrzej Ciarkowski et al., "Methods for Detection and Removal of Parasitic Frequency Modulation in Audio Recordings," in *Audio Engineering Society Conference: 26th International Conference: Audio Forensics in the Digital Age*, 2005, http://www.aes.org/e-lib/browse.cfm?elib=13246.

While I do not foresee the potential of this work to remedy this issue in real-time, the proposed method of inserting an ultrasonic clock in-band, married to the audio, could allow for the realignment of audio that has travelled through disparate latent audio paths.

4 Methodology

Through my professional experience and countless engagements in scenarios involving issues of time-of-arrival, I resolved to explore solutions to the issue. The proposed solution of a high-resolution in-band clocking signal emanating from a source giving purchase for an algorithm to defeat the temporal distortion and ultimately the positional movement of microphones came fairly quickly in these early ruminations. This quickly became the goal; to explore if this approach was feasible. If LTC could be recorded in-band and used to restore the temporality of parallel recordings, surely an ultrasonic signal recorded alongside audible information could lead to similar solutions.

With these early visions of what could be possible, the subsequent work followed this vision in an attempt to realize and implement its potential. I did not begin with a particular methodology in place before exploring the domain of audio production and identifying issues within current practices of the field. A bottom up would approach would start with a methodological foundation, allowing for an emergent yet unknown project to unfold. I saw this neither as an advantage nor disadvantage as, while I sidestepped the challenges of finding my project through exploration, I had the challenge of plotting a course towards a particular frame of desired outcomes. With these beginnings, my explorations of methodology were directed at scaffolding my vision and plotting a course towards my desired outcomes.

This work eschews a formal academic design methodology. This partially due to the top-down discovery of the problem and solution coming before engaging with a formal thesis process. A methodology is an approach. While it would be disingenuous to align this work under the umbrella of a complete design methodology, the following are components of the overall framework used in completing this work.

4.1 Narrative-Driven User-Centred Design

Users play a central role both in the investigational vectors plotted by my research questions as well as my prototype goals of creating a useable time-of-arrival alignment tool. In formulating my initial inquiries, I too was a user that experienced the issues of TOA and could possibility benefit from its alignment. The work of this thesis comes from the observation of a user's need and ultimately aims to engage users with the results of this investigation. As such, this thesis incorporates elements of user centred design (UCD) in its design methodology. Rather than use standard ethnographic approaches within UCD, I have opted to instead implement a narrative inquiry approach. The approach of Gausepohl et al uses a storytelling protocol that collects narratives from both stakeholders and designers[11]. A process of narrative analysis follows where 'user needs are explored and design opportunities are identified.' The narrative process allows for tacit and explicit knowledge to identify these user needs, facilitating a design process that moves from the 'problem space' to the 'solution space'.

This project engages a storytelling protocol similar to that proposed by Gausepohl et al. By interviewing experts in audio production and engineering, I have analyzed individual narratives to synthesize common narratives that illuminate user needs in the problem space, allowing for my design to create solutions to address those needs.

4.2 Research through Design

My thesis work is an example of research-through-design as it endeavours to both create a specific solution applied to the world through design as well as extending knowledge within the field through

[11] A. Gausepohl, Kimberly, Woodrow W. Winchester, Tonya L. Smith-Jackson, Brian M. Kleiner, and James D. Arthur. "A Conceptual Model for the Role of Storytelling in Design: Leveraging Narrative Inquiry in User-Centered Design (UCD)." *Health and Technology* 6, no. 2 (July 2016): 125–36. https://doi.org/10.1007/s12553-015-0123-1.

research.[12] As research, this thesis provides contributions of abstracted theory and, as design, produces a situated realization of a prototype.

It is an important distinction that this thesis is research through design rather than research for design. The design component propels the research component rather than the research scaffolding the design. Through my work, I am probing the issue of time-of-arrival of a source at multiple microphones. My prototype design work is intended to examine that issue. While the lines between research and design blurs as they propel one another forward, it is this kernel of intent that frames how I approach the theoretical writing and prototypical making.

Both frameworks of research for design and research through design are useful as lenses to unpack and reveal what lessons specific designs could have in a larger theoretical context as well the inverse, revealing the potential general abstract theory can have in my specifically situated design solutions.

4.3 Iterative Prototyping

This work endeavours to create a working prototype of a system that will correct for differences in time-of-arrival between microphone sources. In the scope allowable in this thesis, the prototype will exist virtually, moving 'outside the box' only after satisfactory completion of a virtual prototype.

With the ultimate design conceived of before this thesis approach, the process of making the prototype is an iterative exploration and execution of methods and technology in order to reach the ultimate goal of time-of-arrival alignment.

An example of the iterative making is the execution of a clock interpolation engine. The earliest iterations of this implement a crude parallel 'dummy clock' that is a stand in for the eventual in-band

[12] Mads Soegaard and Rikke Friis Dam. 2013. *The Encyclopedia of Human-Computer Interaction*, 2nd Ed. (2nd ed.). The Interaction Design Foundation.

ultrasonic clock. Through multiple successful iterations and countless failures, this algorithm evolved to meet the specific demands of the overarching design goals.

4.4 Learning by Teaching

This work is borne from my years as a professor of media production arts at Ryerson University. In fact, the initial concept was sparked in my intermediate audio class. In this class, students make the leap from recording sources with a single microphone to capturing complex sources with multiple microphones. This exercise focuses on the principles and considerations of combining the perspectives of multiple microphones, highlighting the benefits and challenges of moving beyond a single microphone. It was at this moment that the concept of positionally-aware microphones came to me.

Years later, in engaging in a formal thesis process with a cohort from a wide range of disciplines and experiences, I quickly realized and was informed that my thesis would require attention to not only the intrinsic challenge of time-of-arrival alignment but also to the extrinsic challenge of connecting my work to a larger audience outside the domain audio production professionals. By connecting these ideas with a general audience, it forced me to better connect with the core principles of the problem and its solution, let alone the rationale for solution.

Through this work, I hope to document, illustrate, and present on this topic on an ongoing basis. By connecting the work to others, I continue to explore and refine my own understanding of it.

5 Existing Designs

As this work and resultant prototype will manifest as a professional production tool, the bulk of compatible projects examined fall into that classification. In the world of professional audio production tools, there are a number of designs that reflect elements of this work. There are many tools that finesse the time, phase, and pitch of audio; even tools that work at the same level of precision that this work aims for.

5.1 Celemony Capstan

Autotune ushered in a new era of music production; the pitch of any performer could be tweaked. Cher's Life after Love is the most obvious example of purposefully egregious autotuning, but countless tunes have slight deviations in pitch seamlessly corrected for good or ill. Can't sing? No problem! Want to sound like a warbly robot? Also, not a problem. Antares' Autotune and Celemony's Melodyne are still the leaders of this technology although many newer players have entered the space; the tools have become quite commonplace, available to professionals and novices alike. Antares took an automatic set-and-forget approach to correcting pitch, whereas Melodyne provided advanced editing tools to allow for custom pitch-design of a performance using their patented DNA technology.

In 2012, Celemony directed this technology for correcting vocal pitch towards fixing warped recordings in a software release titled Capstan[13].

Capstan is a commercially available audio processor that corrects wow and flutter pitch deviations present in recordings. It attempts to realign recordings with their original temporal reality and, in doing so, their original rate and pitch. Capstan analyzes the original recording, determining instances within the recording of wow and flutter. The software is looking for global changes in pitch in the recording

[13] "Celemony | Capstan," accessed November 22, 2018, https://www.celemony.com/en/capstan.

that are indicative of a recording or transfer timing issue. With the wow and flutter identified, the recording is varisped at a rate inverse to the original deviation, thereby removing it. Advanced settings and editorial tools allow the user to distinguish between intended pitch deviation (vibrato) and the problematic wow & flutter. Users are also able to finesse and fine-tune the process of correcting for the wow & flutter.

This technology has the capacity to resurrect a wealth of recordings and connect them with a new audience. The quality issues of older technology often limit the connection a performance can make with a contemporary audience not able to hear beyond the effect of wow and flutter. By realigning the recording with real-time, the veil of technological error and perspective is lifted, making a stronger connection between the artist and audience. This is the ultimate conceptual aim of this thesis to make a stronger and more transparent connection between artist and audience. On a technical level, the dynamic, automatic, changes in pitch (varispeed chase) that Capstan provides connects it to the investigations of this thesis.

5.2 Dan Dugan Automixer

The life of a production audio mixer can be stressful business. Take your typical roundtable discussion scenario on the nightly news as an example. It's unscripted, anyone can contribute at any time; sometimes two panelists are talking over one another. Leave all the microphones up in the mix and the mix sounds unfocused and ambient; the mix lacks definition and fidelity. The current speaker is present not only on their own mic; their voice also contributes to everyone else's mic. The sound of their voice arrives at their own, relatively proximate, microphone first before arriving at all other microphones. The other microphones are also more distant with a higher ratio of ambience to voice present in their off-mic contribution. These off-mic qualities fight the speaker's own microphone in the mix, detracting from the quality and precision of the mix.

The Dan Dugan Automixer[14] is a commercial audio tool that automatically mixes multiple microphone scenarios. Using a proprietary process, the Dugan compares the relative level of incoming signals. The strongest signal is passed while the relatively lower signals are attenuated. If one person is talking, they contribute to their own microphone primarily and bleed into other microphones at much lower level. The Dugan allows the highest signal to pass, attenuating all other sources, diminishing them in the mix. If two channels in the system have an active contribution, the Dugan allows both signals to pass, but attenuates them both reduce the level of two people speaking over one another.

Figure 8 - Dan Dugan Automixer for Waves Multirack[15]

The Dugan defines the contemporary epoch of broadcast production audio quality. This simple yet effective process allows for more dynamic and unscripted oral scenarios to be reproduced in a range of mediums. The communication potential of hearing a large freeform group as a whole and as individuals is made possible by the Dugan Automixer. As such, it enlarges the scope of possible forms of engagement between artist and audience while simultaneously making a stronger connection between artist and audience through enhanced clarity.

[14] "Dugan Automatic Microphone Mixers," accessed November 1, 2018, https://www.dandugan.com/products/.
[15] "Dugan Automatic Microphone Mixers."

Dan Dugan's work relates to this thesis in the clarity and focus it brings to complex multi-microphone situations. In a tangential fashion, the Dugan addresses time-of-arrival issues by simply removing multiple microphones from the mixed image, a blunt but effective approach.

5.3 Syncro Arts VocAlign

Issues arise in production audio scenarios that compromise the quality of the recorded audio. Great performance, but an airplane flew overhead. Great delivery, but something was wrong with your microphone. Nice acting, but there's no way we could get useable audio with all the pyrotechnics. Great recording, but your performance was lacking.

Fear not. Ambient, technical, and performance issues can be solved by replacing spoken dialogue with new recordings. ADR is the agreed upon term for replacing production dialogue with dubs. It is standard practice to replace a varying amount of dialogue in a narrative film or television production for either technical or performance reasons.

It's a big ask to get actors to recreate the emotion and action of a scene out-of-the-moment. You're no longer on a ship on wild seas fighting dragons, you're in a comfortably lit sound room with croissants. On top of that, you have to replace lines with the exact same timing as the original. If not, it can easily look like a bad foreign language dub a-la-Godzilla. With beeps in your headphones and wipes on the screen, the recordist tries to help you with your sync as much as possible, but it's still really challenging. ADR is often out of sync with the audio it's trying to replace. Matched with the original visuals, this detracts from the potential connection of the audience to the diegetic narrative world created.

VocAlign[16] is a commercial audio processor that synchronizes replacement dialogue (ADR) to the timing of the recording it replaces creating a perfect match between ADR and the visual. VocAlign scans both

[16] "VocALign PRO 4 - Overview - Synchro Arts," accessed November 18, 2018, https://www.synchroarts.com/products/vocalign-pro/overview.

original production audio and replacement audio. The aural content and cadence of the performance are analyzed and mapped temporally. VocAlign then morphs the replacement audio to match the original.

Figure 9 - VocAlign PRO 4[17]

VocAlign allows performers and artists to make a stronger connection with their audio through the removal of timing inconsistencies between the aural and the visual. Out-of-sync dialogue can be very distracting and detracting from the perceived quality of a piece. Bad sync reveals the apparatus of film making to the audience, breaking down the fourth wall; compromising the narrative contract between artist and audience.

This concept of temporal correction connects with the explorations and intent of this thesis. While the offset and variations of pitch of moving microphones are not of the magnitude that would resonate as a lip sync error, the morphing from one temporal framework to another provides insight into these investigations.

[17] "VocALign PRO 4 - Overview - Synchro Arts."

5.4 Sound Radix Auto Align Post

Sound Radix's Auto Align Post[18] is a post-production audio processing tool that promises to align time-of-arrival between microphone. The process analyzes a recording and processes other recordings to align their time-of-arrival. This process fixes the time-of-arrival of multiple microphones to that of a single microphone.

While the overall objectives of Auto Align Post and this investigation do have parallels, there are key differences in its technical methods and outcomes. Both endeavor to align time-of-arrival across multiple microphones. However, without a reference to temporality from production, Auto Align Post is aligning the time-of-arrival between microphones and not to a designated source.

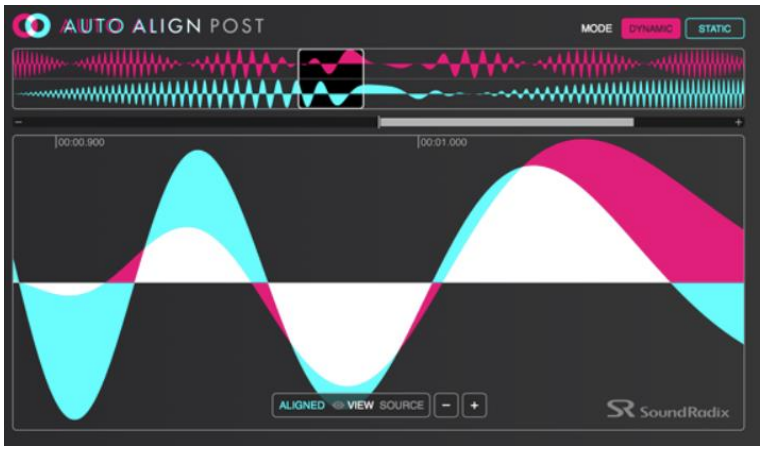

Figure 10- Sound Radix Auto Align Post[19]

Through the proposed methods of this thesis, the prototyped solution will not only align microphones to one another but also to the source. The microphones are aligned and fixed relative to the source. There will be no perceived motion towards or away from a tracked source. In Auto Align Post, if the master recording moves towards away from a source, all microphones will follow suit and move in tandem. The

[18] "Auto-Align Post: The Fast & Simple Way to Get Location Mics in Phase," Sound Radix, accessed November 2, 2018, https://www.soundradix.com/products/auto-align-post/.
[19] "Auto-Align Post."

untracked and uncorrected variable distance between microphones and source is a key differentiator between Auto Align Post and this investigation.

Another key difference is the implemented temporal correction method. Due to its proprietary nature, the exact method implemented by Auto Align Post is unknown. However, as no production temporal recording or metadata is required, the timing and subsequent processing are approximated by an analysis algorithm of unknown design. While this adds a layer of convenience as no extra measures are required in production, there is potential for misinterpretation and misalignment without a production temporal reference.

As the release of Auto Align Post is a recent development, its impact on dialogue mixes and overall soundtrack quality has yet to emerge.

6 2020 Sound

6.1 Conceptual Framework

While the problem of time-of-arrival manifests in post-production, the issue stems from the physical realities of time and space in production or initial recording itself. In the context of this project, production refers to the production scenario wherein sources are captured with microphones.

The composite imaging problem of TOA is the result of relative differences in TOA of each microphone relative to the source. Documenting the relative position of the source to each microphone would provide 'breadcrumbs' that could be used to align the microphones in post production.

The time-of-arrival of a source at each microphone is visually apparent when inspecting the waveforms of their recordings. Their respective time-of-arrival of the source is baked in to the recording itself. If relatively closer, the source is recorded relatively earlier and vice versa. If the source emitted an aural clock signal, that could be used as a reference for the source's time-of-arrival at each microphone.

An audible clock cannot be dismissed out-of-hand. An audible pulse of a single frequency could be removed after the fact using a notch filter to eliminate the frequency of the clock. However, this clock would be a nuisance in production, distracting members of the production, actors delivering lines for example.

An ultrasonic clock above the human audible frequency range of 20Hz to 20,000Hz could be used to silently document the relative location of a source. While we cannot hear anything above 20000 Hz, contemporary audio equipment including microphones and recorders are capable of recording ultrasonic frequencies.

6.2 Production

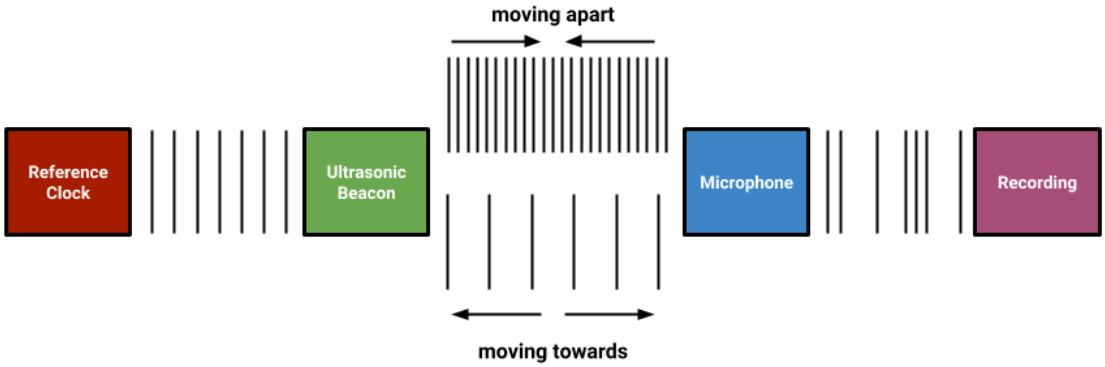

Figure 11- Production Design Overview

In production, an ultrasonic beacon that emits a clock pulse would be placed at the source to be tracked. It is unacceptable to see microphones in narrative film productions, so the same could be assumed of any other technology that is out-of-place in the diegetic space. The beacon would need to be hidden from the camera at a static distance from the source. For the purposes of this project the mouth is the positional reference to the sound source of an actor. Hiding the beacon on the sternum of the actor would provide a relatively static position in relation to the source. This compliments a common placement for lavaliere microphones as, for male and female alike, the sternum often affords a small cavity in which to hide the microphone underneath clothing. Once powered, the beacon constantly emits the reference clock signal. In future designs, beacons would be assigned a unique frequency to allow for multiple beacons to track multiple actors in a scene.

The beacon is the only novel or additional piece of physical hardware demanded by this design. However, in order to transduce and capture the ultrasonic clocking beacon, the microphone and recording apparatus used would have to be able to transduce and record at those frequencies. While on the esoteric end of the microphone spectrum, there are film microphones such as the Schoeps MK41 capsule with CMC6xt microphone amplifier combination that are able capture up to 40kHz. However, while this capsule is fairly common in professional production, the extended range amplifier is not.

Lavaliere microphones are a different story with a common frequency response ceiling of 20kHz, below the ultrasonic range. This problem is compounded by wireless transmission systems used to connect recorders to microphones wirelessly with even the newest generation of wireless transmitters topping out at a ceiling of 20kHz. While this does not bode well for ultrasonic clocking beacon to lavaliere compatibility, it is not a severe setback to the design. The beacon would be placed at or near a lavaliere or, at the very least, in a static position relative to the source. As a result, the lavaliere is already closely aligned to the source. If other microphones, say plant or boom, were aligned to the beacon they would become aligned to the lavaliere as a result.

Contemporary field recorders are designed to record digital at sample rate of 96000kHz, making them compatible with an ultrasonic beacon by facilitating the recording of signals as high as 48000kHz as per the Nyquist-Shannon sampling theorem[20] which dictates that a sampling rate of x can reproduce a maximum frequency of x/2.

Beyond these technical consideration, the production process proceeds as usual albeit with the addition of ultrasonic beacons fixed to sources. This fulfills a goal of the design, to require minimal intervention or disruption in the current workflow of contemporary productions. The design has potential benefits for productions that make use of it, but does not fundamentally change the equipment required or process fulfilled in production. If, for whatever reason, the ultrasonic beacon fails in production, the production audio is still viable in a traditional sense, lacking only the ability to be aligned automatically in the post production component of the design.

[20] "Nyquist–Shannon Sampling Theorem." Wikipedia. March 01, 2019. Accessed March 18, 2019. https://en.wikipedia.org/wiki/Nyquist–Shannon_sampling_theorem.

6.3 Post Production

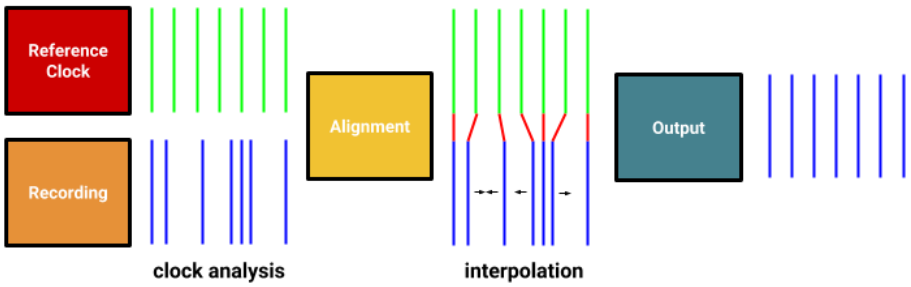

Figure 12 - Post Production Design Overview

The design solution for the post production time-of-arrival tool takes the form of an offline audio processor plug-in. As with most time correction utilities, the very nature of its time manipulation process prevents it from operating in real-time. Ultimately, this would take the form of a Pro Tools Audiosuite, VST, and Audio Units compatible processor allowing for its application in a wide range of digital audio workstation (DAW) environments.

The processor relies on audio recordings with a common temporal and acoustic scenario; that they were recorded in the same time and space. Production recording devices document this information in metadata referring to the shot and take number, even if only in the filename. If implemented correctly in production, the recordings also contain timecode metadata that documents the time of the recording.

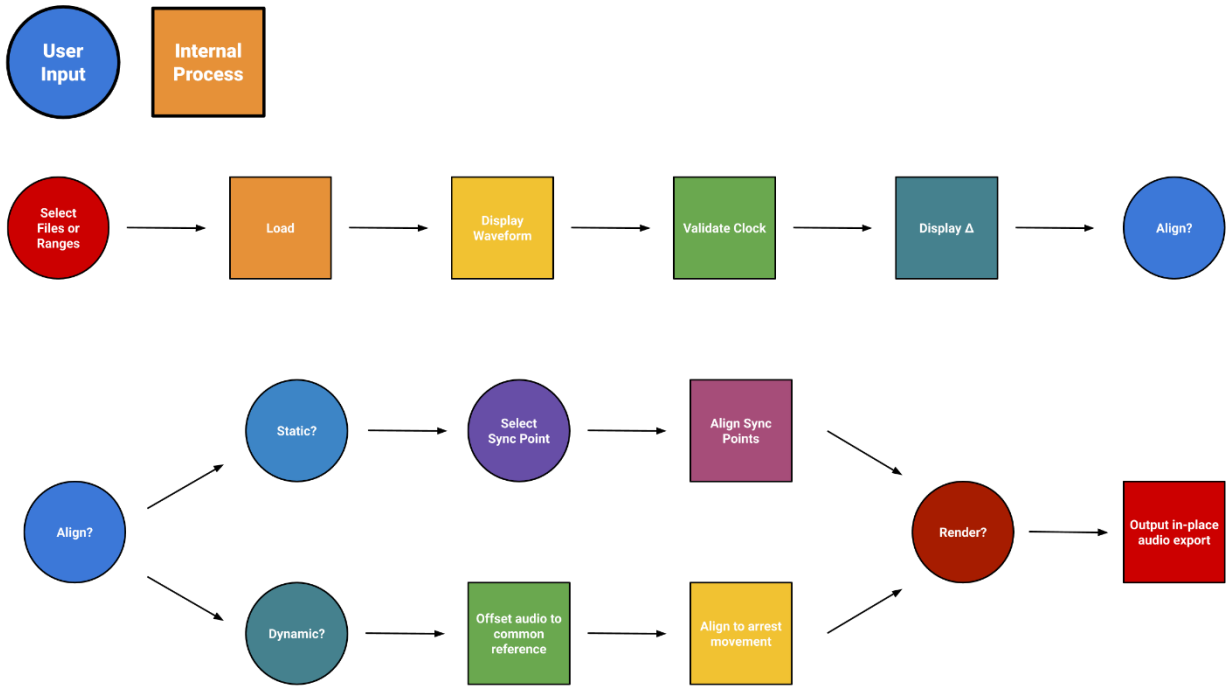

Figure 13 - User and Process Flow

The user selects a set of compatible recordings and loads the DSP time-alignment processor. These recordings are then displayed visually as waveforms. The processor analyses the recording set for valid ultrasonic clock pulses and indicates the results by highlighting the waveforms that comply. Only recordings that have valid ultrasonic clock information will ultimately be processed. As discussed in the production design, lavalieres will not capture the ultrasonic clock signal but are already aligned to the source. The deviation between recordings is displayed in a graph. This demonstrates the magnitude of misalignment between the recordings.

The user can choose to dynamically or statically align the recordings. A static alignment aligns a selected point in time across all recordings, but does not dynamic correct for movement throughout the recordings. The processor defaults to aligning the top of the recordings if no point is specified in static alignment. This choice allows for a simpler, less invasive alignment of recordings, suitable for microphones and sources that do not move during a recording. This is an ideal setting for aligning microphones on a drum kit, for example. This option would be no different than the manual editorial

process. The dynamic alignment process aligns the microphones throughout duration of the recording, virtually negating their movement.

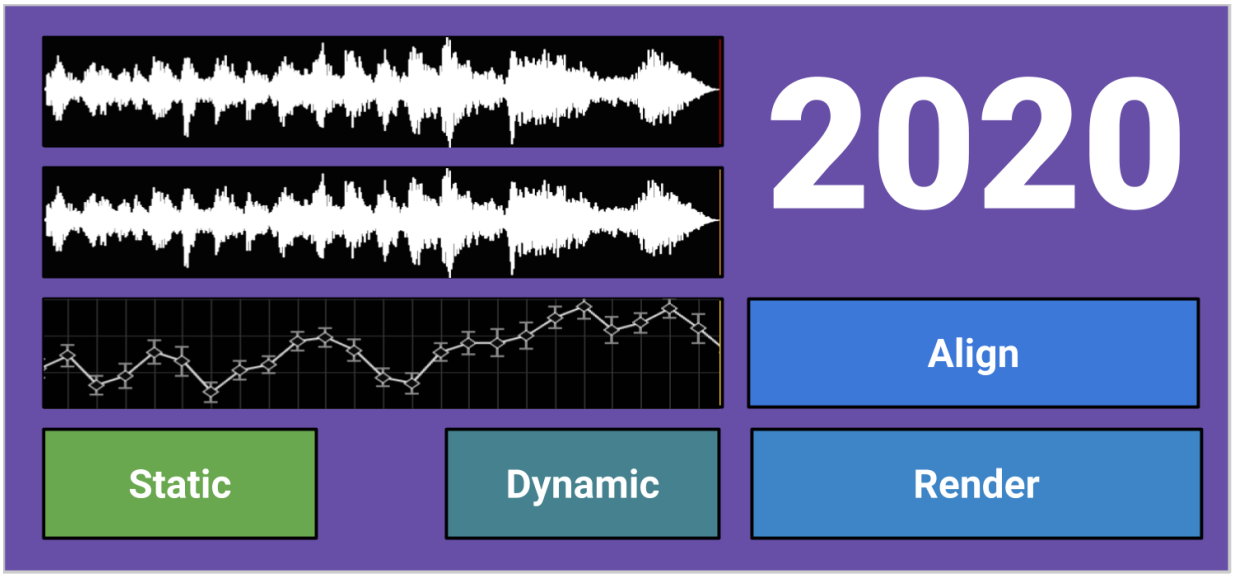

Figure 14 - User Interface

With the alignment type selected, the alignment can be previewed, playing from directly through the interface after the processor computes the alignment. If satisfactory, the alignment can be committed and returned in-place into the DAW environment from whence it came.

The integration will vary depending on the DAW and processor plug-in format but, while the processed alignment will return in-place versions of the original unaligned audio, it will not destructively overwrite or replace the original recordings. Rather, it will create new processed files. This is standard practice for offline audio processors.

The user can proceed and continue working with the processed clips in their DAW as they would the original recordings.

6.4 User Narrative Design

As part of my design process, I sketched out the user experience for my prototype design in narrative form. These narratives are synthesized, based on my own experiences in addition to experts interviewed as part of my narrative research. Creating functional narratives for the intended design was invaluable for identify the key components and user flow of the ultimate design.

6.4.1 Production

Jan is a production sound mixer for narrative & documentary film and television productions. She owns and maintains a recording kit that includes lavaliere, shotgun, and small diaphragm condenser microphones. These microphones track subjects in a scene as body, boom, planted, and boundary microphones. In addition to affixing body mics to her subjects, she discreetly places the 2020 Sound localization beacons to each subject she wants to track. She double checks that each beacon is on and has a unique address assigned to distinguish each beacon. Jan spec'd her microphones, wireless transmitters, and recorder to ensure they can record the ultrasonic clock pulses of the beacons. Jan records ISOs and mixes of each scene as per usual, delivering them to the production for use in offline picture editorial and dialogue editorial.

6.4.2 Post Production

After receiving a locked cut of a film, Krystin begins the dialogue editorial process. As usual, she cuts dialogue from the source ISO recordings to sync with picture and cuts in alternate takes as required. After the production dialogue is edited, patched & finessed, Krystin launches the 2020 sound offline processor. Krystin selects multiple clips that belong to the same temporal recording set. She clicks 'Analyze' and 2020 begins scanning the selected clips for a valid ultrasonic positional clock signal. If valid, Krystin can select which source she wants to align to. For clips with multiple actors, Krystin assigns body mics directly to a beacon. Depending on who's talking at any given point, 2020 dynamically changes its

alignment to focus on the actor currently speaking. Krystin, previews the alignment, tweaking the fine tuning settings as needed. Satisfied with the alignment, Krystin clicks 'Commit', rendering aligned audio to new clips on a new editorial playlist.

7 Prototyping

The prototype explored in this thesis is a virtual proof-of-concept for both an in-band ultrasonic clock for use in production as well as a DSP time-of-arrival alignment algorithm. While this does not progress as far as real microphones in an acoustic space, it forms the basis of the technical underpinnings required to make this design function.

In this context, the audio source is somewhat irrelevant as an existing recording can stand in for a source. The microphone perspectives are affected translations of the source material. By adding a unique initial temporal offset (delay) to each copy of the source, distance relative to the unaffected reference source is created. By uniquely modulating the temporal rate of the files, the movement of microphones throughout the recording is synthesized. While these processes exist virtually outside of an acoustic space, they mimic the offset of perspectives and their movement.

In attempting to align these perspectives back to the reference source (the original temporality), the challenge becomes reversing the processes that make the perspectives unique and dynamically changing. The roadmap to achieving this begins in the ultrasonic clock. The ultrasonic clock is married to the sonic material before the material is cloned, offset, and modulated temporally. It is this clock that creates the 'breadcrumbs' that allow us to find our way back the temporality of the reference audio.

7.1 Ultrasonic Clocking

Generating a reference tone above the audible spectrum proved a simple task. Audacity was able to generate a tone of any frequency allowable by the format of audio file. The sampling rate, the number of amplitude samples recorded per second, dictates the maximum frequency the recording can capture or reproduce. The Nyquist theorem states that a sampling rate of x can reproduce a maximum frequency of x/2. The common sampling rate of 48000Hz therefore can reproduce frequencies up to 24000Hz. While this did leave some room to generate ultrasonic tones between 20000 and 24000 Hz, I

opted to increase the sample rate to 96000Hz to allow for more spectral distance between my clock signal and the audible spectrum.

To start, I was most concerned with the rate of the clock, less so with the position. The clock needed to document the rate of time, but not necessarily what the time is. The clock position at this point is determined by the start of the clock as this becomes a common reference to a single instance across all recordings of the source. In future work, I endeavour to create a clock that both documents rate as well as a continuous refreshing position, such as longitudinal timecode (LTC). For more information on LTC, refer to Appendix A: Methods.

In creating a clock to translate rate, I began by cutting a continuous tone into regular intervals, creating pulses of tone. To start, the cycles of tone and silence were equal in length. This formed a binary of alternating states, tone on and tone off.

It was quickly apparent that, despite creating the clock using inaudible ultrasonic tones, the clock created was indeed audible. The on/off cycles of ultrasonic frequencies created an emergent lower audible frequency.

One of the first clocks I created was using a tone of 25,000Hz. The clock was split between pulses 1000 samples long and silence of equal duration. This duty cycle created a pulse wave that existed very much in the audible spectrum.

These results held true despite changing the frequency of tone, the frequency of the clock, or the on/off duty cycle of the pulses. There was still an emergent audible frequency in each of the clocks.

A similar problem occurs when transforming signals from the time domain to the frequency domain through a fast Fourier transform. A range or window of time is selected to be translated by an FFT. This very selection, however, has an impact on the transformation. The type of window has a frequency

response that impacts the transformation between domains. The ‘windowing’ of my pulses created spectral artifacts just as they would in an FFT.

Through softening the transition from tone to silence, thereby removing the binary of states, the audible frequencies disappeared. While a most welcomed development, the pulses had lost some of their definition. The clock was no longer a binary, but rather ultrasonic tones fading in and out. A downstream challenge would be in parsing a clock signal that did not have a clear edge between states.

The first successful iteration of a clock was now complete, in that it could aurally translate a rate without being audible. This was a workable first clock iteration that would allow exploration of DSP alignment techniques. There would be many iterations to come as I developed the alignment system and gained feedback from the alignment system.

7.2 DSP Alignment System

7.2.1 A Dummy Clock

The exploration of a DSP TOA alignment system began without sound. Ultimately, this system would take clock pulses, measure their rate, and correct for deviations from the reference rate.

As such, these explorations began with methods to analyze a clock signal and correct it. Once I was confident in this system, I would connect it to the acoustically-viable clocking signal.

My development started by creating a very simple clock system that could not be acoustically transmitted. This would act as a proxy for the eventual acoustically-viable clock. In creating arrays with on/off cycles, I created my first rudimentary clock. The first iteration was an array with one positive value for every 100 indexes. The positive value was 1 and silence was zero. This effectively mimicked a very simple clock pulse with a frequency of 1/100 indexes.

Because of its binary nature, this clock was easily analyzed and the distance between each pulse could be easily extracted from the array.

7.2.2 Clock Distortion

With this first clock in hand, a varispeed engine was created to affect the clock signal. Starting with a static varispeed, the rate of the clock pulses within the array was stretched through interpolation. For this process, nearest neighbour interpolation worked quite well, as I was not concerned with the value of the clock, but more the position. By interpolating the array to 110% of its original length, I was left with clock pulses repeating every 110 indexes. However, as each index was stretched, there were now two clock pulses every 110 samples. Using a simple *for loop*, any index that was not the leading edge of a clock pulse was turned off, creating a cleaner array with 1 clock pulse value in every cycle.

In shortening the array, compressing it to 80% of its length for example, some clock pulses disappeared altogether. This was the result of the nearest-neighbour interpolation.

Moving to linear interpolation solved this problem but required a more advanced loop to resolve the binary to the leading edge of the pulse. Only the first non-zero value in a pulse was promoted to full-strength, wherein the rest were demoted to a value of 0.

With the clock data stretched and cleaned up to only document the leading edge of the clock pulses, the analysis of the array to extract the duration between clock pulses was straightforward and predictable. A varispeed of 110% resulted in a 1 clock pulse every 110 indexes.

7.2.3 Alignment Algorithm

The reference clock of 1/100 and stretched clocks provided the building blocks for a crude alignment system. The alignment would come from correcting the stretched clock back to the 1/100 frequency.

Both clocks started in sync, so their misalignment stems from their difference in rate. Correcting the rate would effectively correct their alignment.

With a clock of 1/110, interpolating each cycle duration to 100 indexes would realign the clock with the reference. Using 110 as an example, dividing 110 by the reference value of 100 provides the indexes of interpolation. Resolving these values to an array of 100 samples completes the alignment.

7.2.4 Parallel Processing

This abstract process of distorting and realigning clocks provided a strong enough foundation to implement with real audio examples.

The 1/100 clock was matched to a short audio clip. The clock was repeated so that its length matched the number of audio samples in the clip. These two arrays, clock and audio sample data, remained separate but now matched in terms of length.

The same varispeed process was applied to both clock and audio sample data. Both arrays were temporally distorted by the same amount.

After cleaning up the clock signal to identify the leading edge of the clock sample, the applied distortion and resultant clock cycle length was measured. The measurement of this clock signal was then applied to interpolate the audio data back to the reference clock. The distortions to the clock provided a map to restore the audio data back to its original temporality.

Aiming to improve upon the process, methods of polynomial interpolation were applied to enhance the transparency of the process.

The restoration of temporally distorted audio worked in a static sense. The audio data played back as if nothing happened. It wasn't perfect, playing it back against an inverted copy of the original illustrated

some artifacts of the process. However, this phase check did reveal that there were indeed fewer artifacts in employing polynomial interpolation.

7.2.5 Modulating Temporal Distortion

The next challenge was to employ a modulating temporal distortion, known musically as vibrato, to the audio material and see if the same clocking method would hold up.

This vibrato effect modulates the speed of the audio data over time. Using a sine function, the vibrato method continuously warps the audio at a specified rate and by a specified amount. The matched clock and audio data were subjected to the modulating time distortion and was able to be realigned with the reference audio. Audio processed with this effect sounds warped, alternately increasing and decreasing in pitch. The purpose of this temporal distortion was to probe the capabilities of my alignment methods. They had previously worked on static varisped audio, but it was unknown if it would work with a fluctuating temporal distortion.

While ultimately a developmental placeholder, the modulating pitch of this warped audio is analogous to moving microphones, albeit on a different scale. If a source and destination are moving toward one another, the cycles of compression and rarefaction emanating from the sound source are temporally compressed, resulting in a higher perceived pitch. The inverse is true when source and destination are diverging. The divergence results in expansion of the cycles of compression and rarefaction are lengthened, lowering their frequency. Thus, as microphones and sources are constantly moving their cycles of compression and rarefaction transduced by the microphone are either being compressed (pitched up) or expanded (pitched down) depending on the movement.

The synthesized vibrato created by these methods would act as a proof-of-concept source for aligning microphones. It’s important to note that he audible temporal and pitch distortions of my vibrato method are exaggerations of an order of magnitude beyond the temporal and pitch distortions created

by moving microphones. In reality, we only perceive these distortions on objects travelling at significant speed, the doppler effect of a passing train for example. The movement of microphones in a production context, do not register as doppler as the magnitude of movement of source and destinations is at an order of magnitude lower, therefore not registering as a perceivable doppler effect.

Without modification to the alignment method, I was able to restore the warped audio to its original state. A continuous tone (sine wave) that I had warped was no longer pitching up and down; it was now completely flat. Music examples had the same dramatic results, I was unable to tell the original audio, free from temporal distortions, apart from the distorted and subsequently realigned results. Further phase analysis did reveal artifacts of the process, but these artifacts were not apparent to me the listener without the phase analysis process.

7.2.6 Real Clocks

Propelled forward by the positive realignment of warped audio, the project was ready to move to the next phase. So far, realignment was achieved through parallel temporal distortion of parallel clock and audio data arrays. The clock at this point was not integrated into the audio itself, rather it was a separate parallel entity. The goal of this development is to have the clock integrated into the audio itself, residing in the ultrasonic domain alongside the sonic content of the source. The parallel clock model was an effective proof-of-concept for DSP alignment, but it was unknown if these methods would work with an embedded ultrasonic clock.

The parallel iteration of the clock was simple to analyze. In that work, the clock was a simple binary of values. The current iteration of clock analysis would prove to be more of a challenge as its complexity had evolved to be audio pulses in the ultrasonic domain. Before merging the clock in-band with audio material, I endeavoured to analyze and extract a rate from the ultrasonic clock alone.

My first inclination to detect the clock signal was to use free Fourier transforms to look for activity in the frequency domain to detect the frequency of the clock pulse. While I was indeed able to quantify activity of the clock in the frequency domain, I found it difficult to precisely identify the leading edge of the pulses. When translating from the time domain to the frequency domain, a window of samples is used to create the translation. As the window is the range of samples used in the translation, the results of the translation cannot be used directly to pinpoint a moment within that window. By comparing strengths of the clock frequencies across different windows allowed for a rough approximation of the leading edge, but it was not precise enough to satisfy the requirements for the proposed time alignment application.

Another challenge to this relatively rudimentary FFT approach is the doppler shift in pitch caused by movement of the microphones. The reference clock has a fixed pitch that's easy to identify. However, this pitch shifts a the microphone moves relative to the source, making precise analysis another variable to contend with in an FFT detection model.

7.2.7 Feature Detection

Moving beyond this approach, I explored methods of feature extraction to identify the leading edge of each pulse. Librosa is an audio feature extraction library for the python scripting language and 'provides the building blocks necessary to create music information retrieval systems.'[21] One of Librosa's core features is onset detection, identifying the onset of notes or percussive transients in a piece of music. As the clock pulses are transients, sudden changes in sound pressure level, the intent was that Librosa would be able to detect their onset. However, as the pulses exist in the ultrasonic domain, they were largely ignored by the Librosa onset detection methods. As they were not audible, they were not registered by the system designed to detect audible musical events. However, cheating the sample rate

[21] "LibROSA¶." LibROSA. Accessed March 19, 2019. https://librosa.github.io/librosa/.

of the clock audio to a lower rate, thereby making the pulses audible, allowed Librosa's onset detection to perceive the pulses with some degree of accuracy. The sample rate of the clock data was 96000Hz, allowing for the clock pulses to exist in the ultrasonic spectrum. However, instructing Librosa to process the audio as if it was sampled at 2400Hz dramatically reduced the pitch of the clock pulses, bringing them back into the audible domain and allowing for their detection by Librosa. With this cheat, Librosa interpreted the ultrasonic clock and returned arrays documenting sample positions of the leading edge. The position was off by a number of samples, but consistently so. While off somewhat positionally, the rate of the clock was accurately resolved by the onset detection process.

Lowering the sample rate on input for Librosa's onset detection method resulted in longer durations for interpretation and ultimately used vast amounts of memory that prevented the onset method from executing. Dividing the data into manageable chunks of 100,000 samples allowed the method to function. A method was created to separate the audio data into 100,000 sample blocks. The onset detection results were sequentially appended to arrays and the offset of the block added to the results. This created a transparent process wherein a large audio file could be processed. However, onsets at or near the start of the block were ignored and omitted from the results. As the current working clock had a cycle of 2000 samples, this occurred at every block. As a workaround, albeit a temporary one, the block size was altered to an interval that did not cycle sympathetically with the clock. However, despite this fix there were still some pulses that aligned with the beginning of an onset detection cycle and subsequently missed by the scan. To resolve these gaps in the clock, a method was created to add a clock pulse where one was missing. For example, if a gap between detected pulses is significantly higher than the reference, the missing pulse is interpolated between the two detected pulses.

7.2.8 Clock with Audio

With successful detection of the ultrasonic clock pulse, the next step was to achieve similar results with the ultrasonic clock mixed in-band with standard audio data. The data used was my default musical example. Originally recorded at 44.1kHz, the material had a maximum frequency of 22050Hz resultant from the Nyquist theorem and confirmed with spectral analysis.

The example clip was resampled to 96000Hz making its framerate compatible with the ultrasonic clock. Following this, a method was created to take a single cycle of the ultrasonic clock and repeat it to match the duration of the audio data. The clock and audio data arrays were then summed. In their summation, both sources were attenuated to avoid overmodulation.

With clock and audio data summed together, the resultant audio was exported and aurally inspected to ensure if the clock was indeed inaudible, due its high frequency. The summation had worked, a visual spectral reading confirmed the presence of the clock in the ultrasonic domain which was completely inaudible in the audible spectrum.

To see if the clock function as expected, the next process was to isolate the clock and analyze it for positional and rate information. The audio would need to be separated to prevent the musical transients of the audio data from being detected as the clock.

To isolate the clock from the audible sonic frequencies, the audio data was filtered with a high-pass filter to pass the ultrasonic frequencies and cut the low frequencies. Interestingly, applying a high-pass filter to the ultrasonic clock had the same effect as fading the pulses in and out. By removing all frequencies below 25000Hz with a 6^{th} order Butterworth high-pass filter, the net effect on the single frequency clock pulses was a short fade in and out of the pulses, softening their transition.

Despite this, this process successfully separated the audible from the inaudible. Running this separated audio through the feature extraction algorithms provided the desired interval information, an array of

sample positions of the pulses' leading edge. With the filter softening the pulses, the interpretation was offset slightly. However, the overall distance or duration between pulses was intact.

After successful analysis of the embedded clock on temporally neutral audio, the next test was to distort the temporality of the clocked audio and see if the clock remained useable. Applying vibrato to audio merged with the ultrasonic clock provided the expected results. The resultant distance between pulses expanded and contracted depending on the phase of the vibrato process.

In order to realign the audio back to the reference, the audio had to be fed through the interpolation engine, using the timing information pulled from the ultrasonic clock. The interpolation realignment algorithm had to be redesigned, tailoring it to the cycles of the embedded ultrasonic clock signal, the reference distance between clock pulses.

Following this modification, the audio was passed through the revised interpolation engine. The algorithm aligned the audio back to its original temporality. The vibrato 'wow' was reversed and the musical example is returned to something resembling its original state, no longer speeding up and slowing down.

This correction was not as precise as the earlier parallel clock experimentations. The increased distance between clock pulses ultimately reduces the accuracy of the realignment algorithm. Whereas the parallel clock iteration restored a continuous tone to a single pitch, there are slight fluctuations present in the embedded clock experiments. These are most apparent with a continuous tone and less so with complex music mixes.

8 Conclusions and Future Work

At this stage in its development, 2020 Sound restores the original temporality back to a temporally distorted audio file. This is a successful, albeit virtual, proof-of-concept that negates the movement of microphones in space.

The prototype demonstrates that an in-band clocking signal can be used to both document an audio recording's temporality and provide the roadmap to realign it back to the reference. Regardless of a microphone's movement in space, its original rate can be restored, reversing the minute fluctuations in its pitch.

This work is an important first step to realize the larger design goals of 2020 Sound with the intent of moving forward to create a practical toolset that would increase clarity, focus, and ultimately the impact possible in soundtrack media.

It is important to state that this technology, or any technology for that matter, is not neutral. While this project aims to improve the potential of soundtrack media, it would have an effect on the creative process and the people that make use of it. The intent of this project is to illuminate the potential and rationale for aligning time of arrival in hopes of bringing more people into the practice. It is not the intention of this project to replace individuals with an automated process, but rather to empower a growing community with the development of a new tool and new potentials for the craft.

8.1 Future Work

In the virtual domain, work can be done to finesse the ultrasonic clock embedded alongside the audible material. Minimizing the distance between clock pulses would benefit the accuracy of the realignment algorithm. The clock pulses implemented in the scope of this project proves that this method is functional, but work must be done to demonstrate its viability as an approach.

The alignment algorithm itself is another vector for improvement in future iterations. The alignment is calculated solely between clock pulses. The rate of correction stairsteps from pulse to pulse, with the amount of rate-change varying in each cycle of interpolation. A spline or polynomial interpolation in the change of rate between clock pulses would effectively smooth this, continuously morphing the rate of interpolation between clock pulses. Interpolating the rate of interpolation would create a smooth curve of rate adjustment between clock pulses. This would reduce the artifacts produced by the relatively wide clock pulses applied in the current prototype.

The intention of the 2020 Sound design to align multiple microphone sources to one another demands that these explorations transition from the virtual to the acoustic. Moving into acoustic space will be a daunting endeavour and will present many challenges to the full realization of the design goals.

Emitting the ultrasonic clock pulse acoustically will undoubtedly present obstacles. The highly directional nature of high-frequency sound, for example, will impact the ability to successfully capture the reference clock pulses. By far, the biggest challenge is the potential impact that reflections and reverberations of clock pulses in an acoustic environment will have on their feasibility in a downstream alignment algorithm. The algorithm will have to be finessed to distinguish between direct clock pulses and reflected ones.

As illustrated, the equipment used to transduce, amplify, route, and record the audio in production also adds uncertainty and challenges to future iterations. Microphones and recorders that can accurately capture and document the clock pulses will be a necessity.

The prototype developed focuses on negating movement of microphones but does not address the initial offset of microphones in time and space. Future iterations will align these offsets at the start of the recordings before processing to counteract movement.

Unique ultrasonic beacons, each with signature clock pulses, could allow for multiple sources to be tracked in a scene. As most scenes involve multiple actors, this could be a necessity to the perceived success of future iterations. Variations in pitch, rate, or details in pulse composition could differentiate sources, allowing for multiple clocks to function together in single scenario.

Bibliography

Mads Soegaard and Rikke Friis Dam. 2013. *The Encyclopedia of Human-Computer Interaction*, 2nd Ed. (2nd ed.). The Interaction Design Foundation.

"10MX." Antelope Audio. Accessed November 13, 2018. https://en.antelopeaudio.com/products/10mx/.

"Auto-Align Post: The Fast & Simple Way to Get Location Mics in Phase." Sound Radix. Accessed November 2, 2018. https://www.soundradix.com/products/auto-align-post/.

"Celemony | Capstan." Accessed November 22, 2018. https://www.celemony.com/en/capstan.

"Chirp." *Wikipedia*, November 2, 2018. https://en.wikipedia.org/w/index.php?title=Chirp&oldid=866897234.

Ciarkowski, Andrzej, Andrzej Czyzewski, Marek Dziubinski, Andrzej Kaczmarek, Bozena Kostek, Maciej Kulesza, and Przemyslaw Maziewski. "Methods for Detection and Removal of Parasitic Frequency Modulation in Audio Recordings." In *Audio Engineering Society Conference: 26th International Conference: Audio Forensics in the Digital Age*, 2005. http://www.aes.org/e-lib/browse.cfm?elib=13246.

Davis, Gary, and Ralph Jones. *The Sound Reinforcement Handbook*. 2. ed., 2. printing. Milwaukee, Wis: Hal Leonard, 1990.

"Dugan Automatic Microphone Mixers." Accessed November 1, 2018. https://www.dandugan.com/products/.

"Interpolation Methods in Scipy." Modesto Mas | Blog, October 28, 2015. https://mmas.github.io/interpolation-scipy.

Niemitalo, Olli. "Polynomial Interpolators for High-Quality Resampling of Oversampled Audio." . . *Introduction*, n.d., 60.

Peters, Nils. "Sweet [Re]Production: Developing Sound Spatialization Tools for Musical Applications with Emphasis on Sweet Spot and off-Center Perception," n.d., 305.

Purcell, John. *Dialogue Editing for Motion Pictures : A Guide to the Invisible Art*. Routledge, 2013. https://doi.org/10.4324/9780203784570.

Rumsey, Francis, and Tim McCormick. *Sound and Recording: Applications and Theory*. Oxford, UNITED KINGDOM: Taylor & Francis Group, 2014. http://ebookcentral.proquest.com/lib/ryerson/detail.action?docID=1638630.

"ST 12-1:2014 - SMPTE Standard - Time and Control Code." *ST 12-1:2014*, February 2014, 1–41. https://doi.org/10.5594/SMPTE.ST12-1.2014.

"VocALign PRO 4 - Overview - Synchro Arts." Accessed November 18, 2018. https://www.synchroarts.com/products/vocalign-pro/overview.

Wallaszkovits, Nadja, Tobias Hetzer, and Heinrich Pichler. "Drift and Wow Correction of Analogue Magnetic Tape Recordings in the Analogue Domain Using HF-Bias Signals." In *Audio Engineering Society Convention 136*, 2014. http://www.aes.org/e-lib/browse.cfm?elib=17189.

Zafari, Faheem, Athanasios Gkelias, and Kin K. Leung. "A Survey of Indoor Localization Systems and Technologies." *CoRR* abs/1709.01015 (2017). http://arxiv.org/abs/1709.01015.

A. Gausepohl, Kimberly, Woodrow W. Winchester, Tonya L. Smith-Jackson, Brian M. Kleiner, and James D. Arthur. "A Conceptual Model for the Role of Storytelling in Design: Leveraging Narrative Inquiry in User-Centered Design (UCD)." *Health and Technology* 6, no. 2 (July 2016): 125–36. https://doi.org/10.1007/s12553-015-0123-1.

Appendix A: Methods

The following is an exploration of clocking components that have potential for temporal documentation of a source in acoustic space followed by methods of interpolation and feature extraction that allow for the alignment of moving microphone recordings.

Clock signals

In order to flatten the temporality of a recording, thereby counteracting its movement in space, a clocking signal recorded in production is necessary. This clocking signal provides a reference to the temporality of the production scene, providing data that creates a baseline for a flat temporal rate. As per the solutions to correction wow and flutter, this baseline can then be used to flatten any deviations that occur, in this case from microphones moving relative to a clock emitting at the intended source.

Rate and Position

In production audio, the simplest form of sync is the clapping slate. By recording the clapping slate, a position is documented across all recordings. Cameras document the visible clap whereas an audio recorder documents the audible clap. This allows for their synchronization in the postproduction process through the simple alignment of the clap present on all recordings (audio and visual) to a single point in an editorial timeline.

However, this singular clapping slate fails to define a rate. If the clocks between recorders are running at a different rate, the resultant recordings will drift from one another as one recording will be faster than another. The sampling rates between recorders may vary, but their ultimate connection to time must match if the recordings are to maintain sync.

A clap before a take (top slate) and after (tail slate) can be used to derive a rudimentary rate based on the duration between both claps. However, this fails to account for any clocking deviations that occur

between the reference claps. It would be impractical to clap a slate during the action of a take as the aural and visual slate would impact the intended purpose of the capture itself. However, the more positional references available, the higher the resolution of the rate that can be derived.

It is a caveat that no two clocks, no matter their precision, will run in perfect sync. Two clocks will always drift apart. Only by sharing a common clock can two recorders maintain sync between them. As such, camera and audio equipment are often connected by base system signal clocks such as Tri-Level or Bi-Level video sync in contemporary cameras or DARS (Digital Audio Reference Signal), also known as word clock, in digital audio recorders. These signals are electronic pulses that create a clock with a relatively high resolution. In the case of word clock, the clock has a frequency that matches the recording rate of the device, 48,000 cycles per second for example. The same applies for video sync signals, with the clock aligning to the frame rate of the production, 24 frames per second as an example.

By connecting clocks in production, the resultant recordings will have a shared connection to the original temporality of the scene. However, while devices have a common reference, this says nothing to the quality of the devices' temporal connection to time itself. Rather, the imperfections of the master clock will be translated to all slave devices. While audio equipment can be connected to the output of a atomic clock[22], most practical clocks are less exacting.

LTC

The system level clocks of video and audio systems exists only to define a rate between devices, allowing them to run in sync with one another. The systems fails, however, to translate a positional reference to time. Their frequency describes the rate of time but cannot describe what time it is.

Longitudinal timecode (LTC) allows for the translation of both the rate and position of time. LTC is a Manchester encoded biphase signal that allows SMPTE timecode (Hour, Minute, Second, Frame) to be

[22] "10MX," Antelope Audio, accessed November 21, 2018, https://en.antelopeaudio.com/products/10mx/.

documented with each frame. The phase of each pulse translates bits of data that when decoded describe the position of time. A rate emerges from the continuous burst of positional references.

The frequency of LTC resides in the audible frequency spectrum allowing it to be distributed electronically and recorded by standard audio equipment. While the signal can be leveraged to drive the clock of a recorder it can also be recorded as a standard audio source. In this context, the original temporality of the clock can be restored regardless of the rate or variations in the recorders clock.

Despite its utility in documentation the position and rate of time, the audible nature of LTC makes it unusable to track the relative position of microphones in space.

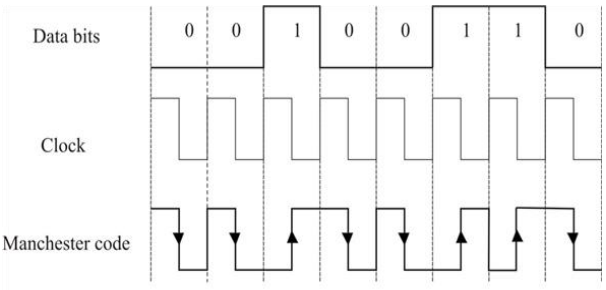

Figure 15- Manchester biphase encoding[23]

Ultrasonic Chirps

A clocking signal above the audible spectrum would allow for such a signal to exist acoustically without being audible. Human hearing tops out at 20,000 Hz but varies from individual to individual. Anything above this 20kHz is above the range of human hearing. While outside of the range of human hearing, a wide range of audio equipment is able to capture above 20,000 Hz successfully. Contemporary professional audio recording equipment is able to record with sampling rates as high as 192,000 samples per second, allowing them to capture frequencies that max out at 96,000 Hz. This allows for a significant bandwidth of recordable sound that is above the human hearing threshold.

[23] "ST 12-1:2014 - SMPTE Standard - Time and Control Code," *ST 12-1:2014*, February 2014, 1–41, https://doi.org/10.5594/SMPTE.ST12-1.2014.

Ultrasonic chirps are a common method for ultrasonic ranging and localization that use short bursts of inaudible ultrasound to transmit a clock pulse. The chirps are defined by their modulation of frequency (FM). The design of the chirp can have the frequency sweep from one value to another over the length of the chirp.

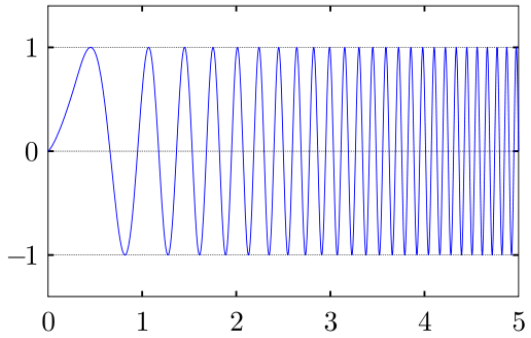

Figure 16- Frequency modulated chirp sweeping up in frequency[24]

Positional localization methods implement these chirps measure the time-of-arrival of a source at multiple perspectives, a microphone array.[25] The inverse is also a valid approach, to measure the time-of-arrival of multiple unique chirping sources at a single microphone. Regardless of approach, the time-difference-of-arrival (TDOA) allows for the calculation of the relative location of an object in space. Whichever microphone has the earliest TOA is closest to the source, whichever has the last is furthest. A one-dimensional array can locate an object's position between microphones. A two-dimensional array can locate an object on a single plane. A three-dimensional array adds height to the equation allowing for the calculation of a location in three-dimensional space.

[24] "Chirp," *Wikipedia*, November 2, 2018, https://en.wikipedia.org/w/index.php?title=Chirp&oldid=866897234.
[25] Faheem Zafari, Athanasios Gkelias, and Kin K. Leung, "A Survey of Indoor Localization Systems and Technologies," *CoRR* abs/1709.01015 (2017), http://arxiv.org/abs/1709.01015.

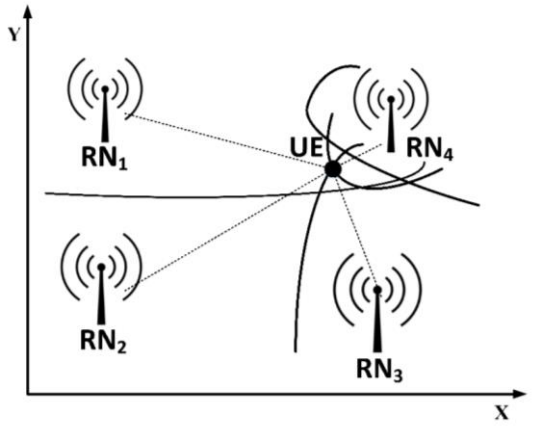

Figure 17 - Time of Arrival positional localization[26]

Interpolation

To nullify the Doppler effect created by movement of either source or microphone, the inverse movement is applied the recordings. As source and microphone converge, slowing the recording has the potential to cancel the recorded movement. Conversely, divergence is counteracted by an increase in playback rate. In order to fix this varisped playback to a new recording, the playback output is fed a recorder operation with a consistent clock speed. As an analog example, the output of a varisped playback machine is connected to the input of a temporally consistent recorder.

Fixing varispeed in a digital context applies the same principles through interpolation. As a digital recording is a sampling of amplitude at fixed intervals of time, techniques of interpolation can be implemented to change the playback rate of audio while maintaining the technical sampling rate of the audio file. Interpolation allows for the creation of new data points between existing data points in a set. For example, through interpolation a playback rate of 46,000 samples per second can be translated to a file with a standard rate of 48,000 samples per second. Through this rate change, new samples are

[26] Zafari, Gkelias, and Leung.

created from the existing data set to effectively fill in the gaps between the existing samples. This allows for a change in playback rate in a digital context.

Zero Order Hold and Nearest Neighbour

The simplest forms of interpolation does not calculate new data points from the existing set. In zero order hold interpolation, an interpolated data point equals the value of the previous fixed data point. Nearest neighbor creates new data points that equal the nearest fixed data, regardless if it falls before or after.

While these techniques do allow for a change in audio playback rate, they create audible artifacts.

Linear Interpolation

Linear interpolation goes beyond these techniques by formulating a straight line between existing data points. Data points interpolated linearly exist on the line drawn between the two points. This dramatically improves the audible quality of interpolated audio data beyond nearest neighbor or zero order hold techniques.

Polynomial Interpolations

Polynomial interpolation considers multiple points (if not all) in a data set to define a curve, or more specifically, a spline between existing points. However, a polynomial that considers too many data points is susceptible to Runge's Phenomenon wherein increasing the number of points calculated through polynomials generates errors in interpolation. Limiting the quantity, or order, of fixed points used to generate interpolated points can avoid this phenomenon.

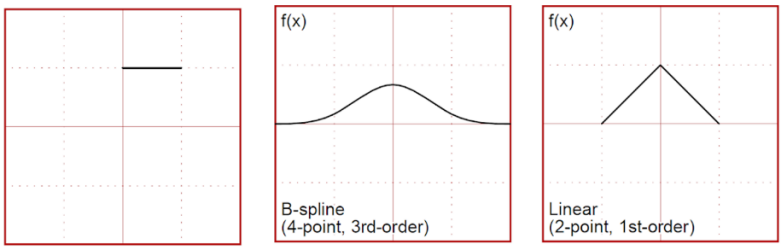

Figure 18 - Zero Order Hold, Linear, Polynomial Interpolation[27]

Modesto Mas's blog post on Python polynomial implementation[28] breaks down polynomial interpolation in Python programming language using the SciPy libraries. SciPy is a library of mathematics, science, and engineering resources for Python which includes a wealth of functions that facilitation a range of methods for interpolating data. Modesto's blog connects the theory of polynomial interpolation with its execution.

Feature Extraction

In the context of this work, audio feature extraction provides purchase to detect sonic clock signals, thereby allowing for their interpolation back to a reference.

In an audio context, feature extraction tools analyze audio in order to identify features. This could be transients, the sudden change in sound pressure level indicative of a percussive hit. A feature extractor with an onset detection algorithm identifies these transients within a recording and reports their positions. Another common feature extracted using these tool is frequency or pitch extraction. Through this, a feature extractor can identify notes in a musical composition.

As an onset detector provides purchase to identify the leading edge of transients, it could be used to determine the leading edge of an ultrasonic clock pulse. However, as to not confuse the algorithm with audible sonic information that may include transients, the ultrasonic clock would need to be filtered

[27] Olli Niemitalo, "Polynomial Interpolators for High-Quality Resampling of Oversampled Audio," . . *Introduction*, n.d., 60.
[28] "Interpolation Methods in Scipy," Modesto Mas | Blog, October 28, 2015, https://mmas.github.io/interpolation-scipy.

from the rest of the audio spectrum. This filtering isolates the clock from the rest of the spectrum, in hopes of only leaving the clock intact for detection.