Allegheny College
Department of Computer Science

Senior Thesis Proposal

Sound of Words

by

Enpu You

ALLEGHENY COLLEGE
COMPUTER SCIENCE

Project Supervisor: **Douglas Luman**
Co-Supervisor: **Douglas Jurs**

October 19, 2020

Abstract

Sound of Words

Building on the understanding that speech has unique sonic signatures, this thesis presents a novel way of using computational strategies to represent the musical qualities in languages. Sounds of Words explores the polarization in ideology and social discourse in the current world through active listening, transforming sounds of human speech into a musical soundscape.

Sound of Words proposes a system that can automatically extract musical elements such as pitch and duration from live human speech to create music compositions. The project strives to adopt an automated computational strategy to explore musicality in the nature of human speech, to experiment with the existed music concepts, and to draw attention to a long-neglected perspective of the human listening process. It is an attempt to integrate music and computer science in an intelligent and artistic way. The application will be used to present a live music/art performace and the source files will be fully open-source available.

Introduction

Context

Humans often forget speech has sonic features as well as music does. I first noticed this when watching an NBA highlight on my phone — Lebron James blocking Andre Iguodala in the game seven of the finals. The commentary "Block by James!" came out of the speaker. A friend of mine who happened to have perfect pitch was sitting next to me with his cello. I noticed he was pulling the bow and some weird noises came out. But soon, I realized it is the exact pitch contour of the commentary I was watching.

Music is widely believed to be related to language in various ways, as people always say "music is a universal language." Regardless of the purpose of music or language, we can agree they both share a lot of similar vocabularies when we are describing them, such as antecedent, consequence, phrase, syntax, accent, pitch, rhythm, meter, tone, and so on. Computer scientists, therefore, were inspired to use Natural Language Processing (NLP), a technique used for machines to understand natural language like English, to construct music. James R. Meehan, professor at the Department of Information and Computer Science at the University of California, Irvine, suggests that using NLP allows researchers to look at both the harmony (local syntax structure) and also the phrase (higher-level semantic structure) in analyzing and generating music[18]. Human ears are intentionally used, most of the time, for either language or music, however, human brains are more used to responding to the semantic contents of language and the sonic features of music. We tend to forget that language(speech) itself contains musical elements like pitches as well. If we take language as a composition of words, then speech is its performance. Nonetheless, only upon repetition of a word or phrase do human start to capture these musical patterns with the brain. Studies suggest that there is even a relation between the widely accepted tone combinations (major and minor scale) and human speech. Neurobiology researchers from Duke University claimed that excited speech contains more major intervals while subdued speech contains more minor ones[16]. In general, there are just a lot of details we have been neglecting and taking for granted in speech.

However, sonic features are not the only things we have overlooked, speech itself is neglected all the time. It is part of human nature to tend to surround ourselves with others who share the similar perspectives and opinions about the world[12]. Due to the rise of digital media and social media platforms, such as Facebook and Twitter, and their personalized recommendation algorithm, we often get stuck in these "echo chambers" where certain speech and discourse have been repeated, amplified, and reinforced while others have been simply excluded[12]. At a higher level, it leads to the increasing social and political polarization. These polarizations happen both in political parties as well as the citizens. According to a 2014 study from the Pew Research Center, partisan ideology has been more divided than any point in the last two decades[8]. Around 35 percent of the registered party members view the other party as a "Threat to the Nation's Well-Being"[8]. In 2014, 92 percent of Republicans are to the right of the median Democrat, while 94 percent of Democrats are to the left of the median Republican, which is over 30 percent higher than two decades ago[8].

Objectives

This project, *Sound of words*, aims to implement a natural language processing (NLP) tool to extract the sonic features from a variety of contexts of human speech (e.g. self-talk, speeches, conversation, et al.) by utilizing technologies like audio analysis and speech recognition. Sonic features might include but not limit to pitch, rhythm, duration, amplitude, etc. Secondly, it attempts to further explore the relation between music and speech, closely examine the basic music concepts such as melody, duration, and rhythm, and discover the inherent musical qualities of speech by curating a live sound environment for the audience to become composer, performer, and listener simultaneously through their speech. Lastly, the project attempts to represent social discourse through sound and promotes the idea of active deep listening.

Related Work

As I have looked through sources from various disciplines, there has not been a project quite similar to what is being proposed here: to collect sound elements from human speech for music composition. However, a lot of inspirations for this project are from researches in various areas such as music, computer science, cognitive psychology, neuroscience, and linguistics. Thus, in this section, I will be introducing some important related researches in these areas.

Music and Language

When we use the phrase "Music is a universal language," we mean that we can perceive the emotion conveyed in music without a certain written syntax or semantic content. However, we tend to neglect that language/speech also has melodies, rhythms, and all the other musical elements that music has when focusing on its semantic content. And thus, we are always subconsciously perceiving these hidden features in speech as they are being performed. Academically, music and language have been researched more commonly as two separate disciplines over the past decades[22]. Linguists would study speech and language, while musicologists would mostly focus on music. More combined and comprehensive studies have been done about music-language interface in recent years, surprisingly, by cognitive psychologists and neuroscientists for exploring the origin of music and language and how the brain processes sound.

Cognitive psychology aspect

Music, language, and the brain is claimed to be the first comprehensive study of the relationship between music and language from the standpoint of cognitive neuroscience from 2007 by Aniruddh Patel, a professor of psychology at Tufts University. Scholars believe that Patel has challenged the belief that music and language should be processed independently, and the "music-language relations have barely begun to be explored"[22]. The book provides a great amount of evidence of music and language's connections, in categories such as pitch, rhythm, melody, meaning, and so on. As Patel points out, the comparative research done on music and language is either on the differences or the commonalities. Most of the work has focused on the former. Thus, the book not only provides an extensive discussion on the structure of musical and linguistic systems but, more importantly, an alternative perspective to emphasize the commonalities over differences. The cognitive psychology side of this study offers an insight into how music and languages are similarly processed. It has proposed methods to extract statistical regularities from rhythmic and melodic sequences, which will greatly assist the discussion in this project about a closer examination of music and language, and further designs on how language can be used to produce music. According to Patel's another work on rhythm and melody of speech, musicologists and linguists have suggested that the prosody of a culture's native language is reflected in the rhythms and melodies of its instrumental music[23]. Its result shows that music reflects patterns of durational contrast and pitch interval variability between successive vowels in spoken sentences between British English and French. This 2006 research was conducted with audio files and the manual transcriptions of them.

Music, Language, Speech and Brain: Proceedings of an International Symposium at the Wenner-Gren Center is another comprehensive source that includes a collection of articles related to speech/language and music from the conference *Music Language Speech and Brain.* The book provides broad overviews of research in the various fields related to these topics in order to promote the general recognition of this interdisciplinary field. Six main aspects were presented: descriptions of language and music, speech and music performance, voice and instruments, cognitive and perceptual aspects, neurophysiological aspects, principles applied when speech and music are combined[28].

Computer science aspect

Music-language interface A recent study from 2018 investigates similarities and differences between poetic and musical meter and melody[19]. The group utilizes a system, Munich Automatic Segmentation System (MAUS), to automatically annotate the poem audio rendition based on its syllabic units. These annotation grids are then imported into a phonetic software application PRAAT. The corrections were operated manually, for example, marking silent periods larger than 200 ms as pauses[19]. With this approach, the group annotates the pitch of syllables with wave frequencies, then further visualizes them into pitch contours. Mean pitch values and durations, inspired by Patel's 2006 work, are transformed into semitones and musical note duration using the MIDI convention[23]. Although the main topic of this research, poetic melody, might not be relevant to this project, its sound pre-processing strategy and feature extraction and transformation could be an inspiration when implementing the interface for this project. The sound features of human speech would be similar to that of the poem renditions (pitch contours and rhythms) and MIDI is a commonly used machine-readable format.

Algorithm-generated music and NLP technique in music analysis Bruce York's 2019 master thesis on the status of Artificial Intelligence in music provided a condensed timeline for AI in music generation[31]. The first AI music system was developed in 1956 in the approach of Agent. They adopted a Markov process (a random process whose future probability of each event depends only on the state attained in the previous event) to identify the probability of different musical elements following one another[31]. This later becomes a major approach in AI and AI music generation: analyzing existing music, establishing rules, and then generating similar music. A logic approach was also tried in 1958[31]. Two chemistry professors generated a string quartet, *Illiac Suite*, with four sections. Notes were randomly selected into the composition based on counterpoint rules. They generated a cantus firmus and some four-part counterpoints using the Monte Carlo theory (the computational algorithm that relies on repeated randomness to solve problems that might be deterministic in principle) and Markov process[31]. Some musical parameters such as rhythm, dynamics, and timbre were also taken into account in the system. A lot of other works are done in the atonal style in the 1960s, either generating twelve-tone style music, or developing a structural description of atonal music based on Schoenberg's music[31].

Cognitive approaches were also adopted in some cases[31]. Inspired by psychology studies on linguistic theories and how people identified patterns, researchers implemented pattern descriptions and applied techniques of linguistic theory to music generation, and harmony was believed to be the best musical element for the application of linguistic theories[31]. In 1968, Terry Winograd wrote a program that labels chords according to tonal function using the music of Bach and Schubert as the model[31]. Other cognitive approaches believed rhythm is a universal focus of people from different cultures, or generated music from a music "grammar", including rules for meters, chords, and melodies[31].

James R. Meehan has discussed about AI composition based upon total music theory in the journal paper *An Artificial Intelligence Approach to Tonal Music Theory.* Meehan points out that there was very little computer research on models of tonal music theory at the time when this paper was written (1980), as most of the AI composition was done on atonal style music because the composition theory behind twelve-tone music is much easier to simulate by machines[18]. The challenging part for the AI system to generate tonal music is that it relies on a theory that not only analyzes harmony but also gives the relation between harmonic progressions to generate music. Believing that there are strong parallels between music and natural language[18], Meehan and other researchers at the time started to looking into the theory that can represent the linear structure of tonal music: Schenkerian Analysis[1].

[1] Schenkerian analysis, or Schenker analysis, is a model for linear analysis of music[20]. It focuses on the production of graphic representations of hierarchically organized streams. It generally views harmony and voice-leading in three levels: background, middleground, and foreground. As a currently dominant theory of tonal music, Meehan calls it "a transformational system for music analysis" that views melodic motion (horizontal) as a temporal expansion of a harmonic structure (vertical). The transformations "reduce groups of notes on one level to single notes on the next higher level"[18]. The higher-level notes are said to be "prolonged" by the lower-level notes. Nonetheless, Schenker analysis is still considered to be imprecise, and is even a subject of debate by many music theorists.

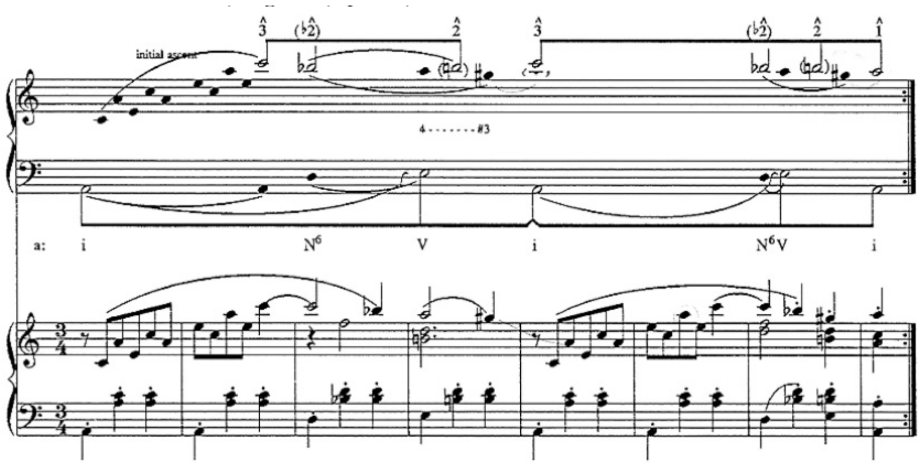

Figure 1: an example of Schenker analysis on Beethoven, Bagatelle, op. 119, no. 9 [20]

The revision made by Eugene Narmour to Schenker's theory in 1977 proposed a new way to look at music: "prediction with absolute certainty"[18], meaning to make predictions from the existing phrases. Meehan suggests that using Natural Language Processing (NLP), a technique used to understand language like English, allows researchers to look at both the harmony (local syntax structure) and also the phrase (higher-level semantic structure) when analyzing and generating music[18]. An optimal direction is to produce a harmony of a given melody in the style of the chorale.

Google Bach Doodle is an experiment of this approach using Bach chorale for training. In the blog post *Coconet: the ML Model behind Today's Bach Doodle* and the research paper *Counterpoint by Convolution*, the researchers behind the Bach Doodle explained their machine learning model that was trained on 306 chorale harmonizations by Bach. They take a piece from Bach, randomly erase some notes, and ask the model to guess the missing notes from context. This allows the model to start anywhere in the music and develop the material in any order[13].

The implication-realization theory proposed by Narmour is substantially discussed in Ramon Lopez de Mantaras and Josep Lluis Arcos's journal paper *AI and Music: From Composition to Expressive Performance*. They restate an intuition shared by many people — appreciating music has to do with expectation: "What we have heard builds expectations on what is to come"[10]. It further explains the theory proposed by Narmour is based on a set of basic grouping structures, which characterizes patterns of melodic implication perceived/recognized by listeners as motifs[10]. Based on this theory, they designed the AI music performance system SaxEx which performs an automatic parsing of the melody to extract these group structures in its process to "generate high-quality human-like monophonic performances of jazz ballades based on examples of human performers"[10]. It then replaces the inexpressive input notes with the expressive ones with similar properties(the basic grouping structure) that stored in the system.

EMI (short for Experiments in Musical Intelligence) is a program that takes the works of a given composer and produces new works in the style, developed by David Cope at the University of California, Santa Cruz. It is based on a database of source works from real composers; EMI would deconstruct and recombine them back together into new works adopting strategies like voice-hooking and grammar from the Natural Language Processing system SPEAC [14]. David Cope designed the SPEAC system that categorizes chords into “Statements, Preparations, Extensions, Antecedents, and Consequents”[14], based upon the music theory rules and tension of chords.

Music Composition

Composition with human speech

It is rare to find academic sources about music composition with human speech. However, making music from human speech is not a new thing. A podcast from New Sounds showed a collection of music composed from human speech from the last fifty years with different strategies. Some are created by instrumental harmonization of human speech contours, such as "Zero Initiative"[26], a work based on a conversation recorded outside of a music venue by Olga Bell. Some are composed by editing, splicing, looping conversation in layers with music instrument accompaniments, such as "John Somebody" by Scott Johnson[26]. An even earlier piece from the mid-60s, "It's Gonna Rain", by Steve Reich, used two tape recorders to playback the voice of a preacher man saying "It's gonna rain" in an out-of-sync pattern[26]. It is clear that some of these music composed of human speech is harder to listen to from the standpoint of the type of music we are used to. However, there is also a recently popular artist on YouTube, Charles Cornell, who makes jazz harmonization to human conversations (mostly celebrity meme videos) went viral[9]. Cornell essentially picks out the contour of a speech and then plays a jazz improvised harmonization behind, which is also in sync with the tempo and phrasing of the speech to create a funny effect. These two sources have provided great insights into various ways of music composition with human speech and will be useful for the composition algorithm design in this project.

Sound art

John Luther Adams John Luther Adams is believed to be one of the most original musical thinkers of this century[25]. His works are mostly inspired by the natural world, especially scenes in Alaska. The goal is to "hear the unheard, to transpose the music that is just beyond the reach of our ears into audible vibrations."[24] He believes that each geographical feature and location has its sonic signature. The idea of transposing these features into music is to "tune the landscape."[25] John Adams created a sound/light installation, "The Place Where You Go to Listen," at the University of Alaska. The project composes upon data concerns earthquakes and any activity going through the ground[3]. It is sent from hundreds of seismological, meteorological, and geomagnetic stations. The output is a sound environment, described as a "vibrantly colored field of electronic sound."[25] It is a live musical ecosystem that changes in real-time based on the incoming data streams representing seismic vibrations of the earth. As John Adams puts it:" I thought that it might be a piece that could be realized at any location on the earth and that each location would have its unique sonic signature. That idea—tuning the whole world—stayed with me for a long time."[32]

"The Place Where You Go to Listen" provides an innovative way of using scientific data to create music and sonify geographical features. As the New Yorker article describes, the audience will notice a dense and organ-like sonority that follows the contour of the harmonic series — the "Day Choir."[25] The range of harmonies represents the brightness of the weather if the sun is out or overcast. A "darker, moodier"[25] harmony takes over when the sun goes down — the "Night Choir."[25] The rhythm patterns in the bass are inspired by small earthquakes and other seismic events — the "Earth Drums."[25] There is also a "shimmering sound"[25] in a high register tied to the fluctuations in the magnetic field — the "Aurora Bells."[25]

The project itself is fascinating as a sound installation. As a "musical counterparts to the natural world,"[25] some audiences describe it as a piece of ambient music. On one hand, it is random, real, and unpredictable in a sense that it is purely based on the input data streams of seismic activities, "the humor of the earth"[25]; on the other hand, the rules of transposing data into music is carefully curated by the composer. The decisions made into composition design deeply reflects the composer's inspiration from the natural world and preference of music style. The New Yorker article describes it as a "forbiddingly complex creation that contains a probably unresolvable philosophical contradiction."[25]

Deep Listening | Pauline Oliveros Pauline Oliveros is an American composer who has been a pioneer of electronic music. A lot of her musical works are created as a soundscape and environment for listening, "Sonic environment, "[21] to ask listeners to be simultaneously an audience, a performer, and a composer, a combined responsibility. In 1988, she and two other musicians, trombonist Stuart Dempster and vocalist Panaiotis, descended into a resonant disused cistern in Port Townsend, Washington, where they later recorded the improvisation — "Deep Listening."[21] It is an idea to not only pay attention to musical elements in the performance such as melody, harmony, and rhythm but also to carefully engage with the acoustic space of the performance.

Pauline Oliveros explores the difference between hearing and listening. She argues listening as a lifetime experience that builds upon accumulated experiences where one continuously develop consensual agreements on the "interpretation of sound waves" such as language[2]. The sound wave captured through ears is delivered to the brain where "listen" happens[2]. She believes that hearing is a physical and objective action merely about what happens in the ear, which can be measured by scientists; Listening is a psychological and subjective perception, to give attention to what has been captured by ear. Hearing turns the vibrations in a certain range of frequency into perceptible sounds, while listening compares the sound to experience and interprets its meaning, which could cost milliseconds to years. The listening itself is an act of composing, which is to direct one's attention to what is heard and to make a decision of action[2].

Pauline Oliveros explains Deep Listening as a way of "listening in every possible way to everything possible to hear."[2] She believes that this type of intense listening applies to various areas, including the sounds of daily life, nature, one's thoughts, as well as musical sounds. More importantly, it is a process of learning. It requires the "temporary suspension of judgment"[2] and a willingness to receive new information, whether it is pleasant or not. Listening and learning can also happen at multiple levels, whether it is within oneself, between individuals, or between one group and another (intrapersonal, interpersonal, and group)[4]. The Deep Listening itself is designed to inspire trained and untrained performers to practice the art of listening[21].

In Oliveros' work "Sonic Meditations" from 1971, she explained her goals to "expand consciousness" and for "humanitarian purposes."[21] With the focus on one's inner world, Her work, as activism, met second-wave feminism's appeal that "the personal is political."[21] The meditation project is to create an atmosphere of opening for all sounds to be heard, to approach the world with ears open, and to listen deeply.

Listening through human ears as well as machines

Obviously, there is a difference between listening through human ears and through machine/algorithm. Human ears are susceptible to melody, harmony, rhythm as we have a long history of listening, while machines are more susceptible to the numerical values such as frequencies and audio features such as MFCC that will be explained more in depth in the next section. Hutson analyzed the definition and the constitution of human creativity in calculation, complexity, and meaning. He argues that the "heart of the mystery of creativity"[14] is to understand the meaning of the creation. As computer would not understand the meaning of a complex fugue, humans can understand the difference between a high pitch and a low pitch note[14]. From the standpoint of cognitive psychology, Hutson points out that humans' cognition of musical meaning bases on real-world physical experience, and computer naturally has this fundamental disadvantage. He quoted Marvin Minsky's talk interview with Otto Laske:

> "A machine that was really competent to listen to nineteenth century classical chamber music might well need some knowledge-understanding of human social affairs — about aggression and conciliation, sorrow and joy, and family, friendship and strangership. And clearly, an important aspect of 'understanding' music experience is the listener's experience of apprehensions, gratifications, suspense, tensions, anxieties, and reliefs-feelings very suggestive of pains and joys, insecurities and reassurances, dreads and reveries, and so forth.[14][29]"

Thus, it would be interesting to integrate machine listening with human listening, in terms of the musical features we can intentionally process and the sonic features we subconsciously take granted for.

Activist art

Activist art commonly takes place in the form of image and visualization[1]. However, Ultra-red has turned the focus to ear and proposed to not only understand music as a process of organizing sound but also to find ways that combine the political, popular educational method with the methods of experimental music. They brand themselves as a sound art and popular education collective committed to the practice of listening as a form of organizing. Through their sound project, they express the acoustics of spaces of dissent and the demands and desires in people's voices and silences. In their mission statement, Ultra-red proposes a "political-aesthetic"[1] model that instead of reating aesthetic forms that contain political content, finds the aesthetic forms from political content. They utilize sound-based research and engage with the "organizing" and analyses of political struggles and social movements, such as anti-racism, community development, and the politics of HIV/AIDS[1].

Although this project will not take the form of "Activist art", it will still be most likely established upon a political polarization context. It does not protest towards any ideology but urges humans to involve with active listening. Thus, it would benefit from certain works and strategies from the field of activist art.

Audio Processing and Music Programming

A recent journal paper from Ryerson University on audio signal feature extraction states that an audio signal processing algorithm involves the analysis of signal and extraction of its properties. The audio signals being analyzed are music, speech, and environmental sounds[27]. Speech is the sound produced by human beings. Its frequency range from 100Hz to 17kHz[27]. Musical sounds are produced by musical instruments. It could be classified into a genre, mood, etc. Its frequency range from 40 Hz to 19.5 kHz[27]. Environmental sounds are essentially any sound other than the previously mentioned two. Different from speech and music, it is hard to find periodicity in environmental sounds. There is a trend in the last few years to integrate audio processing techniques with modern machine learning algorithms[27]. Since the performance of a machine learning model depends essentially on the features it has been trained on, signal feature extraction becomes one of the most crucial steps of a machine learning process.

Automatic Speech Recognition

Automatic Speech Recognition (ASR) is a technology enable and improve the human-human and human-machine communication[33]. Fields such as Speech and Speaker recognition have been an intensive research area for decades. Many technologies have been developed and adopted in achieving the goal of automatic speech recognition, such as Gaussian mixture models, hidden Markov models, and the more modern deep neural networks. According to Dong Yu's research on ASR, before 2010, the advancement in the research and application of ASR was relatively slower and less exciting. In the past few years, however, there has been a surge in the demands of speech technology, especially in deep learning[33]. Its application has been essential in removing barriers between human-human interactions, such as speech-to-speech translation systems; it also greatly improved human-machine communication, such as voice search and virtual assistant. Nonetheless, most ASR systems rely on four major components: signal processing and feature extraction, acoustic model, language model, and hypothesis search[33].

Although this project will not focus on developing or utilizing full features of a speech/speaker recognition system, it would likely make use of some open-source ASR software to transcribe the text and integrate speaker identification if needed or desired for the later stage of music composition. It also takes great inspiration from signal processing and feature extraction. It is also worth researching in the future about how to efficiently take live input signals and enhance its audio speech quality.

Feature extraction

Feature extraction means to highlight the most dominating and discriminating characteristics of a signal[27]. There are various types of features that could be extracted from an audio signal depending on its application. The type of feature extraction can be subcategorized into the time domain, frequency domain, joint time-frequency domain, and deep features. This project will further determine what exact features to extract for music composition. For now, I will briefly introduce some of the basic features that are likely to be used for composition.

Mel Frequency Cepstral Coefficients (MFCCs) are derived from the cepstral representation of an audio clip[27]. It represents a short-time power spectrum of an audio clip and mimics the human auditory system closely, thus making it widely used in pitch detection and speech recognition.

Rhythm-based features show the recurrence of patterns over time, mostly found in music and speech (poetry)[27], including speech duration, articulation rate, phoneme duration, pulse metric, etc. Pulse metric is commonly used in speech/music discrimination and music genre/instrument classification.

Fundamental Frequency F0 is the first peak of the local normalized spectro-temporal auto-correction function[27]. It is one of the most used features in tonality based features. For music, F0 is the musical pitch of a note that is perceived as the lowest partial present. in addition to that, there are also Pitch Profile, Harmonicity, and etc.

Amplitude-based features include Amplitude descriptor (AD) and Attach, Decay, Sustain, Release (ADSR) envelop[27]. AD differentiate between sounds based on its energy and amplitudes. ADSR feature is not achievable for most real-time sounds because the decay and sustain are not clearly present. However, Attach and Release are used in timbre analysis.

Deep features are features extracted from the hidden layers of deep learning models. Deep learning models take low-level information (such as MFCC) as input and output high-level features. These features have been used in speaker recognition and emotion recognition[27].

Composing with programming language

A common strategy for composing music from data is through program languages, also known as algorithmic composition. Music programming is also a common form of music production using electronic devices and software, especially in contemporary music. `SuperCollider` is one of the most widely used platforms for audio synthesis and algorithmic composition built in `C++`. It is an open-source software used by musicians, artists, and researchers. It consists of three major components: `scsynth` the real-time audio server, `sclang` the programming language, and `scide` the editor for `sclang`. `sclang` is an object-oriented and functional language that is very similar to `C` and `JavaScript`. It supports interactive programming and live coding, which allows musicians and artists to easily perform digital signal processing[30]. The following code segment shows how to modulate a sine frequency and a noise amplitude with another sine whose frequency depends on the horizontal mouse pointer position in `SuperCollider`[30].

Listing 1: code example from SuperCollider

```
{
        var x = SinOsc.ar(MouseX.kr(1, 100));
        SinOsc.ar(300 * x + 800, 0, 0.1)
        +
        PinkNoise.ar(0.1 * x + 0.1)
}.play;
```

Method of Approach

The project being proposed here is a comprehensive work including techniques in both computer science and music field, with the inspirations of studies from other areas as the early section indicated. The project can be broken down to a couple of main steps: sound collection, feature extraction & preprocessing, and composition. The methods, algorithm, and tools used to implemented this project will be discussed here. As this project is still in its early stage of planning, this section will be work-in-progress and continually updated.

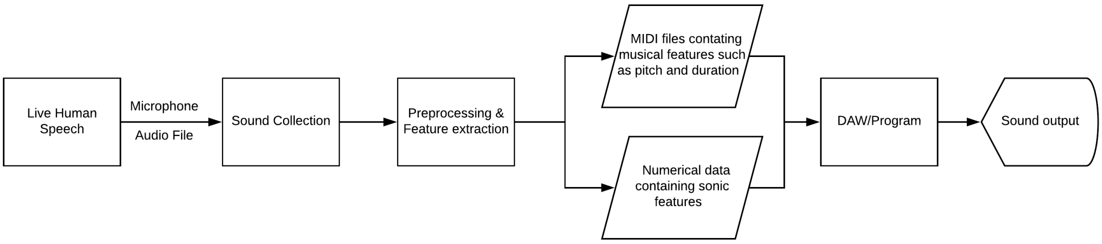
Live Human Speech
Microphone
Audio File
Sound Collection
Preprocessing & Feature extraction
MIDI files contating musical features such as pitch and duration
Numerical data containing sonic features
DAW/Program
Sound output

Figure 2: Flowchart of proposed pipeline

Input and collection

The audio input should first and foremost be human speech. However, it does not matter if the speech is collected from a live environment or through the internet, as it will be determined based on the campus situation by the time this project is presented. The general standard is to be collected in a `wav` file and to contain speech, which could be from live news broadcasting, live speech collected through the microphone, or even a Zoom session. It is acceptable that part of the audio will be pre-determined as a background layer. The ideal state is to also integrate live, real-time speeches so that even though it is random, unpredictable, and relying on unexpected events, the composition will still be organized under certain rules and reflects the personal interpretation. Nonetheless, these should be collected in the same file format and standard which will be regardlessly compatible with the following feature extraction step.

Preprocessing and feature extraction

The second step is centered on the extraction of musical elements such as pitch, duration, volume, and rhythm, as these data will be essential for the composition. For now, I will be modeling this project after the approach taken by the research group of poetic music melody[19]. In their research, they have utilized a tool called `Praat`[6], developed by Paul Boersma and David Weenink from the University of Amsterdam. It is a free open source software for speech analysis in phonetics and continually being updated since its earliest version in 1995. It is available under GNU license on Github with an extensive amount of tutorials provided by its users around the world. `Praat` can be used to do a wide range of analyses, including spectral analysis, pitch analysis, formant analysis, intensity analysis, and manipulation. It is compatible with all the common operating systems, both in its own interface and terminal. In this project, I plan to have the sound files `wav` collected from the previous step input to `Praat` or any equivalent audio analysis library in order to extract the features introduced earlier. A way of implementation is to integrate `Praat`'s terminal functionality into the pipeline of collecting and preprocessing sound files. Another way, which is also more like, is to `Parselmouth`[15], a Python interface of `Praat`, so that the codebase could be more testable and maintainable. Similarly, the audio files should be executed in parallel in the same fashion with other audio analysis libraries to extract additional features that would be helpful for the composition.

It is an essential step to further process the features extracted from the last step in order to make it applicable for music composition purposes, regardless of the possibility of different music composition design. It is expected that the music composition will be generated either through a digital audio workstation (DAW) or a programming language such as `SuperCollider`. Thus, it is important to convert the sound materials into MIDI format, which is a technical standard adopted to play, edit, and record music on electronic musical instruments and computers. Nonetheless, it will be further investigated whether to use a DAW or a programming language to perform composition, depends on the complexity of implementation. Passing data from Python to DAW might introduce new challenges to this project, as to implement a plugin that will let

DAWs such as Ableton to communicate with a Python interface, although there are already some libraries implemented such as `PyLive`.

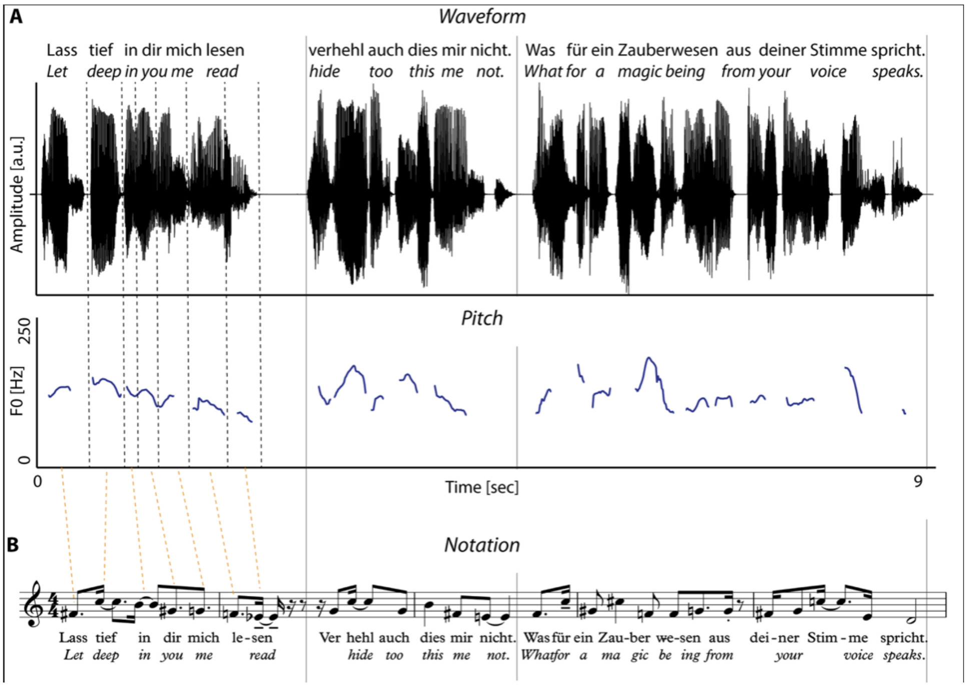

Figure 3: Praat used in poetic melody research

Inspired by the work from poetic music melody group where they did a music analysis on poem renditions[19], raw pitch values (in Hz) will be transformed into semitones on the MIDI scales with numeric values ranging from 21 to 108, using the following formula[19]:

$$d = 69 + 12\log_2\left(\frac{f}{440Hz}\right)$$

with d = MIDI pitch value and f = raw pitch (in Hz).

Poetic melody project also proposed a way to map the syllable durations onto musical note duration, by a simplified assumption that a whole note corresponds to 1s[19]. They chose 16th note to be the smallest duration value, which would corresponded to a minimal syllable duration of 62.5 ms[19]. Although I might not adopt this approach in this project, it could be one of the reference solutions in the future to determine the rhythm and time signature for further composition design. Another approach is to utilize the beat tracking and onset recognition algorithms built inside some other libraries to determine the rhythm and pattern of the composition.

Speech/Speaker Recognition

Depending on the design of composition and the need for text, this project might also incorporate speech and speaker recognition in order to get a transcript of text with the speaker identity in addition to the previous audio features. There are various of speech recognition tools on Github such as pyAudio[11] and speech_recognition[34]. Again, depending on the design of composition, these tools can be input with the audio file generated from the first step and produce text, speaker labels, and other audio features to facilitate the composition.

Libraries

Madmom is an audio signal processing library written in Python with a strong focus on music information retrieval (MIR) tasks. It is easy to use. It allows rapid prototyping of signal processing workflows with no dependencies on other software or machine learning model[5].

aubio is a tool designed for the extraction of annotations from audio signals. It provides features such as onset detection, pitch detection, beat and tempo tracking, and producing midi streams from live audio. It is written in C and has a Python interface[7].

librosa is a Python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems[17].

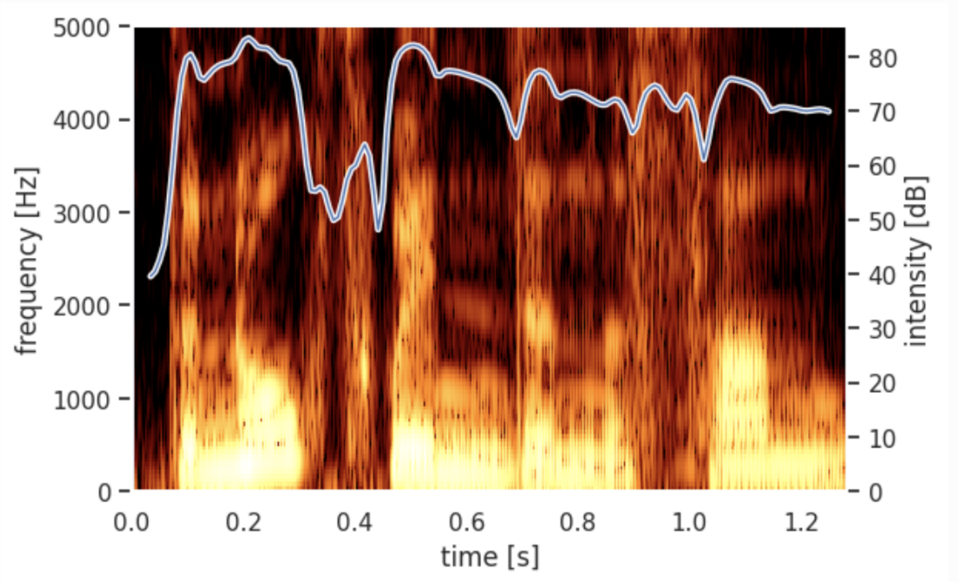

Figure 4: Sample spectrogram generated by Parselmouth

Parselmouth is a Python library for the Praat software. With a pythonic interface, It directly accesses Praat's C code and provides efficient access to program's data[15].

Listing 2: Parselmouth code example

```
intensity = snd.to_intensity()
spectrogram = snd.to_spectrogram()
plt.figure()
draw_spectrogram(spectrogram)
```

```
plt.twinx()
draw_intensity(intensity)
plt.xlim([snd.xmin, snd.xmax])
plt.show()
```

Composition

As previously mentioned, part of the collected sound features would be converted into MIDI format and composed upon through a digital audio workstation (DAW). Many DAW platforms are available to use. Other features could be passed directly into a program that is written in languages such as `SuperCollider`. Although it has not been determined, the project would most likely be using Ableton Live. It is a DAW designed to be an instrument for live music performance as well as a tool for composing, recording, and arranging. Surprisingly, Ableton is also compatible with Python scripts through some open-source frameworks, such as `PyLive`. It would potentially allow the composition to be fully automated. There will also be additional research conducted to explore ways of meaningfully interpreting sound features extracted from the previous audio files and expressing them using different algorithms and instruments based on a set of specific rules.

In addition to the composition that is being generated through algorithms and DAWs, there will also likely to be a part of live improvisation based on the ambient sound environment created by the pre-determined algorithm rules. The improvisation can also be facilitated by certain outputs from the feature extraction step such as key and pitch information. I plan to have an outdoor live event where several artists/musicians can perform simultaneously in the form of a quartet perhaps, including instruments like cello, violin, marimba, etc.

Threats to Validity

A couple of threats to the performace of the system could be noise in the backgroun when collecting input signals. The quality of the signal might also depend on the distance between the source of sound and the microphone. In addition, the quality of the signal would also be subject to the clarity of the collection equipment (i.e microphones). Another risk is that whether the aforementioned speech recognition and audio processing tools can handle conversations with multiple human participating. These could all lead to inaccurate or false results from the sound collection to the feature extraction step.

Evaluation Strategy

Code Testing

There will be at least half of the project implemented in Python code. Thus, code coverage, a measure used to describe the degree to which the source code of a program is executed when a test suite runs, will be a major metric to evaluate the functionality of the tool. A program with high test coverage, measured as a percentage, suggests it has a lower chance of containing undetected software bugs compared to a program with low test coverage. Thus, each python function will be provided with test cases using `Pytest`, a framework that allows building simple and scalable unit tests in python. Code coverage should cover the majority of functions that take input and generate output, and a coverage that is greater than 90 percent will be the goal. In addition, there should be some functional testing to evaluate the performance of the tool under different environments (with background noice) and with different amount of speakers. Since the project also strives to achieve a real-time generation of music, some additional evaluation can be done on the execution time (efficiency) of some major steps in the tool from audio collection to output composition.

Compostion

It is difficult to give a clear evaluation standard for the composition especially at the beginning stage of the project. In general, the final output of composition should showcase the understanding of music concepts and theories. In addition, it is important to have a clear discussion over the choices that have been made throughout the design and composition process in the final thesis, which will mostly likely represent the artistic and musical style of the project.

Research Schedule

Task	**Begin Date**	**End Date**
Proposal defense preparation	Mid Oct.	Early Nov.
Thesis outline	Early Nov.	Mid Nov.
Thesis intro/related works	Mid Nov.	Late Nov.
Sound collection implementation	Early Nov.	Late Nov.
Feature extraction implementation	Mid Nov.	Mid Dec.
Composition	Early Dec.	Mid Mar.
Thesis chpater writing	Late Jan.	Late Mar.
Overall testing	Late Feb.	Mid Feb.
Performance	Late Mar.	Mid Apr.
Thesis defense preparation	Early Apr.	Mid Apr.

Conclusion

Sound of words is a complex project, the implementation and design of composition will not be fully known until initial testing of ideas. The proposed project will implement a natural language processing (NLP) tool to extract the sonic features from a variety of contexts of human speech by utilizing technologies like audio analysis and speech recognition. Sonic features might include but not limit to pitch, rhythm, duration, amplitude, etc. The future work will include further investigation in audio processing and available features to be extracted and incorporated into composition design. In addition, there will be an extensive research into sonic rhetoric in order to better understand the function of sounds in speech.

Bibliography

[1] Ultra-red: Mission statement.

[2] *The difference between hearing and listening | Pauline Oliveros | TEDxIndianapolis.* Nov 2015.

[3] Revamping the sound installation "the place where you go to listen": Alaska earthquake center, Oct 2016.

[4] Anonymous. Deep listening, Oct 2015.

[5] Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F., and Widmer, G. madmom: a new python audio and music signal processing library. In *Proceedings of the 24th ACM International Conference on Multimedia* (Amsterdam, The Netherlands, 10 2016), pp. 1174–1178.

[6] Boersma, P., and Weenink, D. Praat: doing phonetics by computer (version 5.1.13), 2009.

[7] Brossier, P., Tintamar, Müller, E., Philippsen, N., Seaver, T., Fritz, H., cyclopsian, Alexander, S., Williams, J., Cowgill, J., and Cruz, A. aubio/aubio: 0.4.9, Feb. 2019.

[8] Center, P. R. Political polarization in the american public. *Ann Rev Polit Sci* (2014).

[9] Cornell, C. How to turn cardi b into meme music, Jul 2019.

[10] de Mantaras, R. L., and Arcos, J. L. Ai and music: From composition to expressive performance. *AI Magazine 23*, 3 (Sep. 2002), 43.

[11] Giannakopoulos, T. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one 10*, 12 (2015).

[12] Gillani, N., Yuan, A., Saveski, M., Vosoughi, S., and Roy, D. Me, my echo chamber, and i: introspection on social media polarization. In *Proceedings of the 2018 World Wide Web Conference* (2018), pp. 823–831.

[13] Huang, C.-Z. A., Cooijmans, T., Roberts, A., Courville, A., and Eck, D. Counterpoint by convolution, 2019.

[14] Hutson, M. *Artificial intelligence and musical creativity: computing Beethoven's tenth.* PhD thesis, Massachusetts Institute of Technology, 2003.

[15] Jadoul, Y., Thompson, B., and de Boer, B. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics 71* (2018), 1–15.

[16] Kamraan Z. Gill, D. P. A biological rationale for musical scales. *PLoS ONE* (2009).

[17] McFee, B., Lostanlen, V., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., Zalkow, F., Malek, A., Dana, Lee, K., Nieto, O., Mason, J., Ellis, D., Battenberg, E., Seyfarth, S., Yamamoto, R., Choi, K., viktorandreevichmorozov, Moore, J., Bittner, R., Hidaka, S., Wei, Z., nullmightybofo, Hereñú, D., Stöter, F.-R., Friesch, P., Weiss, A., Vollrath, M., and Kim, T. librosa/librosa: 0.8.0, July 2020.

[18] Meehan, J. R. An artificial intelligence approach to tonal music theory. *Computer Music Journal 4* (1980).

[19] Menninghaus, W., Wagner, V., Knoop, C. A., and Scharinger, M. Poetic speech melody: A crucial link between music and language. *PloS one 13*, 11 (2018).

[20] Neumeyer, D. Guide to schenkerian analysis, Nov 2018.

[21] O'Brien, K. Listening as activism: The "sonic meditations" of pauline oliveros. *The New Yorker* (Dec 2019).

[22] Patel, A. D. *Music, language, and the brain.* Oxford university press, 2010.

[23] Patel, A. D., Iversen, J. R., and Rosenberg, J. C. Comparing the rhythm and melody of speech and music: The case of british english and french. *The Journal of the Acoustical Society of America 119*, 5 (2006), 3034–3047.

[24] Poole, N., and Adams, J. L. Always rising: Art and activism with john luther adams, 2017.

[25] Ross, A. Song of the earth.

[26] Schaefer, J. New sounds, Dec 2019.

[27] Sharma, G., Umapathy, K., and Krishnan, S. Trends in audio signal feature extraction methods. *Applied Acoustics 158* (2020), 107020.

[28] Sundberg, J., Nord, L., and Carlson, R. *Music, Language, Speech and Brain: Proceedings of an International Symposium at the Wenner-Gren Center, Stockholm, 5–8 September 1990.* Macmillan International Higher Education, 2016.

[29] Vantomme, J. *Computer Music Journal 18*, 1 (1994), 81–83.

[30] Wilson, S., Cottle, D., and Collins, N. *The SuperCollider Book*. The MIT Press, 2011.

[31] York, B. The status of ai in music: A study of the musical metacreation conferences, 2012-2018. Master's thesis, Western Illinois University, 2019.

[32] Young, G., and Adams, J. L. Sonic geography of the arctic, 1998.

[33] Yu, D., and Deng, L. *Automatic Speech Recognition.*

[34] Zhang, A. Speech recognition, 2017.