Pilot Study Investigating Shared Acoustic Features in Speech and Music Preference

Zain Souweidane

Department of Performing Arts Technology, University of Michigan PAT 421: Advanced Psychoacoustics

Dr. Sile O'Modhrain

December 15th, 2022

Abstract

The theory that language and music perception have the ability to influence each other is being continuously supported by studies across many disciplines. Studies have shown that there are shared acoustic features that encode emotions in speech and in music. This pilot study builds on these ideas by investigating whether features in a person's speech are correlated with the acoustic features in their preferred music by representing each in the same feature space. To do this, speech recordings from seven participants reading a provided elicitation passage were collected. The participants were then asked to like 10-25 second samples from 30 songs. Correlations between acoustic features in the participants' speech recordings and the songs they liked were then calculated. No significant correlations were found between features in speech and preferred music. Due to the small sample size, no substantial conclusions can be drawn from this pilot study.

1: Introduction

Sounds heard in music are usually not directly heard in the natural world. If the exact sound of every musical phrase could be found in common experience, then it could be hypothesized that the reason music creates emotion is because it reminds us of those experiences. However, the sounds heard and the emotions triggered by music are usually more general. Music tends to evoke a wide range of complex emotions that do not always fall into basic emotional categories (van der Schyff & Schiavio, 2017; Vuilleumier & Trost, 2015). This lack of a direct mapping from natural sounds to our perception of music has led to centuries of investigations exploring what might give rise to this phenomenon. One experience that has been looked at when trying to

gain a better understanding of our perception of music is the experience of language and vocal expression.

Music and vocal expression are some of the most common auditory events that trigger emotions. For this reason, many studies have been conducted to investigate shared acoustic features between the two as well as the influence they have on each other. As detailed in the next section, these studies have demonstrated that music perception and vocal emotional perception are heavily correlated and have the potential to influence each other. This pilot study investigates this relationship by observing acoustic features in a person's voice and acoustic features in songs they like. Specifically, the goal of this pilot study was to see if there are acoustic features in a participant's speech that correlate with the acoustic features in their preferred music.

2: Related work

This pilot study was inspired by two fields of research relating music perception and speech perception. One field investigates this relationship by observing the potentially influential relationship language and music perception have on each other. This has been achieved by studying cross-cultural variations in vocal emotional perception and music perception. The other field of research investigates this relationship by analyzing the shared acoustic features that encode emotion in speech and emotion in music. This field has rapidly grown over the past 20 years due to the use of advanced statistical methods to investigate these features.

2.1: Cross-Cultural Variations in Vocal Emotional Perception and Music Perception

Cross-cultural variations in vocal emotional perception and music perception have been studied in order to look at the influence language has on music perception and the influence music perception has on vocal emotional perception. Studies have found that language potentially has

the ability to influence our music perception from a young age, sometimes as early as prenatally. This has famously been shown by studies that found babies' cry melodies are shaped by the ambient spoken language they are exposed to while still in the womb (Wermke et al., 2016).

Another important example of the influence of language on musical perception is the tritone paradox. It occurs when two tones that are related by a half-octave (or tritone) are presented in succession and the tones are constructed in such a way that their pitch classes (C, C♯, D, etc.) are clearly defined but their octave placement is ambiguous. Some listeners hear the sequence of tones as descending while others hear it as ascending tones. Previous studies have established that the perception of the tone pairs is correlated with the listener's language or dialect. It was also found that the first language a listener learns heavily influences the way they perceive the pair of sequences, even if the listener is not fluent in this language (Deutsch et al., 2004). This highlights the potential ability that language has to influence music perception, and that the formation of music and speech perception is critically influenced by experiences early in life.

These bodies of work are some of the many examples that have shown that language and music perception are significantly correlated and appear to have a bidirectional influential relationship. The next section reviews studies that investigate the shared acoustic features that encode emotion in music and in vocal expression.

2.2: Shared Emotional Encoding of Acoustic Features in Speech and in Music

Since verbal communication and music perception have been found to have the ability to influence each other, it would make sense that they might share similar acoustic features.

Furthermore, a certain set of combinations of acoustic features seen in verbal communicative phrases should evoke similar emotions as musical phrases that contain a similar set of features.

By the early 2000s the investigation of emotional responses to features in music and the emotional responses to features in speech were being researched but largely independent of each other. In 2003, a review of 104 studies of vocal expression and 41 studies of music performance was conducted and qualitatively related features found in music to features found in speech (Juslin & Laukka, 2003). Table 1 displays the speech features and musical features that this review related.

Acoustic Feature in Speech	**Acoustic Feature in Music**
Speech Rate	Tempo
Intensity	Sound Level
Intensity Variability	Sound Level Variability
High-Frequency Energy	High-Frequency Energy
F0 (Fundamental Frequency)	Pitch Level
F0 Variability	Pitch Variability
Change in F0	Pitch Contour
Speed of Voice Onset	Tone Attacks

Table 1. Qualitatively related acoustic features in speech and music from (Juslin & Laukka, 2003)

Based on these features, Juslin et. al. found similar results between studies that tested emotional response to music and studies that tested emotional responses to speech. Studies which had speech phrases that evoked anger were categorized by a fast speech rate, high voice intensity, much voice intensity, much high-frequency energy, high F0, much F0 variability, rising F0, and fast voice onsets. Studies that had musical phrases which evoked anger were categorized by fast tempo, high sound level, much sound level variability, much high-frequency energy, high pitch level, much pitch variability, rising pitch contour, and fast tone attacks.

These feature comparisons between speech and music became much easier to investigate as technological capabilities increased over the past 20 years. Specifically, studies have been conducted which leveraged advanced computational statistical methods that predict the emotional sentiment of an audio source such as music or speech. Machine learning methods such as support vector machine models and neural networks are well equipped to identify many features in data that humans might not be able to distinguish. The explosion of applied machine learning over the past decade has helped expose acoustic features unique to music and speech.

Results from an influential study helped provide some of the first quantitative results that support the theory that emotional cues in speech and music are largely encoded in the same acoustic features (Weninger et al., 2013). This study used support vector machine models to find relationships between features in music and speech passages and their labeled valence and

arousal levels. The models were trained on speech data and tested on speech data, trained on speech data and tested on music data, trained on music data and tested on music data, and trained on music data and tested on speech data. The speech and music were represented in the same space using the ComParE Feature Set (Schuller et al., 2013). This feature set contains 6373 features - 65 acoustic low level descriptors (LLDs) and their statistical functionals. When narrowing the ComParE Feature Set from 6373 features to the 200 features that showed the highest correlation across both domains (music and speech), cross-domain valence achieved significant correlations with $r=0.60$ for valence when trained on speech and tested on music. A similar study at the University of Michigan confirmed the results that features from the ComParE Feature Set can be used to represent valence and arousal in both domains (Zhang et al., 2015).

Studies have previously shown that vocal emotional perception and music perception have the ability to influence each other. Studies have also been able to show that music and vocal expression use similar features to encode emotion. This pilot study builds on these ideas by investigating whether features in a person's speech are correlated with the acoustic features in their preferred music by representing each in the space created by the ComParE Feature Set.

3: Experiment

The question this pilot study set out to investigate was as follows: is there a correlation between acoustic features in a person's speech patterns and the music they like? Based on the aforementioned studies, I hypothesized that since language and music perception share acoustic features and have the ability to influence each other then there are acoustic features in a person's speech that will correlate with the acoustic features in their preferred music.

3.1: Study design

The study had participants complete a questionnaire detailing their background and a questionnaire gauging their mood. Participants also provided an audio recording reading a provided elicitation paragraph. After uploading their recording, participants completed the same questionnaire gauging their mood a second time. Participants then listened to 10-25 second snippets of 30 songs and liked or disliked them. Since each participant read the same paragraph, the independent variable can be seen as the features in a participant's voice and the dependent variables are the features in the songs they liked.

The background questionnaire asked questions to gauge participants' general background as well as questions related to their experience with language. The mood survey was used to gauge their current mood. Since this experiment is investigating the relationship between language and music perception, the mood survey was used to establish that song preference was not solely influenced by current mood. The mood survey was taken before reading the passage and after reading the passage. This was used to ensure reading the passage did not affect their current mood.

3.2: Participants

Overall, this pilot study had seven participants. Age, ethnicity, and sex were asked in the background survey as they influence vocal expression. All of the participants were in the age group of 18-24. Five participants were male and two participants were female. For ethnicity, five participants answered white, one answered latino or hispanic, and one answered middle eastern.

As well as basic demographic questions, questions about where they have lived and their experience with language were asked. Since this study is investigating the influence of language and music perception on each other, it was important to know what languages participants speak and have been heavily exposed to. Given that early experiences with language influence music perception, it was also important to know what ages the participants were when they lived in different regions. Participants were asked to note the administrative division they lived in as well as the country. This was to track regional dialects of the same language. Given the small number of participants, this was not helpful but could be useful with a larger sample. All participants currently live and grew up in the United States. One participant lived in Colombia until they were 2 and then moved to the United States. Another participant lived in Syria until they were 9 and then moved to the United States. All the participants spoke fluent English. One participant spoke fluent Spanish as well.

3.3 Apparatus and Stimuli

The mood survey was a survey based on the circumplex model of affect (Posner et al., 2005). Participants checked all the boxes that corresponded with their current mood. Mood options were excited, happy, content, relaxed, peaceful, calm, sleepy, bored, sad, nervous, angry, and annoyed.

The elicitation paragraph chosen for the participants to read was, "*Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station*." This paragraph is used by the Speech Accent Archive and was designed by linguists to accentuate vocal features (Weinberger, 2015).

The 30 songs chosen for the listeners to like were chosen by scanning through random genres on everynoise.com (McDonald, n.d.). This allowed for a wide variety of genres to be drawn from. Songs chosen for the sample set were not well known (based on the number of times they were streamed on Spotify). This was to avoid participants from hearing songs they already knew and had associations with. This way biases in the songs participants liked could be avoided. Once the songs were chosen, they were reduced to snippets ranging between 10-25 seconds. The length of the snippet was determined such that the sample contained a reasonable musical phrase from the song.

3.4: Procedure

The experiment was conducted via an online survey using Google Forms. The full survey can be found here: https://forms.gle/3GouhunkuwUKQGa67. When participants took the survey, they were first asked to upload a consent form. They then filled out the background questionnaire followed by the mood questionnaire. After that they were directed by the survey to record themself reading the elicitation paragraph in a quiet space and upload the audio file. Once they uploaded the file, they were once again asked to fill out the mood questionnaire. Finally, they were directed to the following google folder containing the 30 song snippets: https://drive.google.com/drive/folders/1KkeA0BO8gczfOcSeCLKqOQ2CN5DBOdeQ?usp=share_link. They were then asked to listen to each snippet with headphones and check the box next to the song sample name if they would "listen to it on their own given the opportunity." They were also asked to check the box next to the song sample name if they recognized the song snippet.

4: Results

I hypothesized that since language and music perception share acoustic features and have the ability to influence each other then there are acoustic features in a person's speech that will correlate with the acoustic features in their preferred music. To investigate this hypothesis, the main data sets collected were speech samples of participants reading a provided elicitation paragraph and lists of songs they liked from a set of 30 songs. The songs and speech samples were then represented in the same space using the ComParE Feature Set. The current mood of the participants was also collected in order to isolate another factor that could be influencing song preference.

4.1: Analysis

To extract the 6737 features (65 acoustic low level descriptors (LLDs) and their statistical functionals) detailed by the ComParE Feature Set from each speech sample and song sample, OpenSmile was used (Eyben et al., 2010). In order to find the correlations between features in participants' speech samples and features in song samples a participant liked, a speech matrix and song matrix were created.

For each participant and for every song they liked, the features of that song were added as a row to the song matrix. Likewise, everytime the features of a song they liked were added as a row to the song matrix, the features of their speech recording were added as a row to the speech matrix. This way every row in the song matrix containing the features of a song a participant liked had a corresponding row in the speech matrix containing the participants speech recording. This resulted in two matrices of size 102 x 6373. 102 corresponds to the total number of liked songs across all participants (note the song sample set contained 30 songs so song features are repeated

across rows). 6373 corresponds to the number of acoustic features analyzed. Once these matrices were set up, Pearson's correlation coefficient was calculated between each column in the two matrices. This resulted in a 6373x6373 cross correlation matrix.

As well as the correlation between features in participants' speech and music preference, two other sets of data were analyzed in order to determine the variance in song choice. The number of different songs each participant liked was calculated. The distribution of participants' mood for songs they liked was also calculated.

All the code used to analyze the data can be found here: https://github.com/zainsouwei/LanguageMusicPerception.

4.2 Results

Before analyzing the acoustic feature data, I wanted to establish that the songs participants liked were not completely dependent on their current mood. Figure 1 displays the mood of a participant when listening to a song they liked. Participants' mood state at the time of liking each song was enumerated. As presented in Figure1, it is seen that each liked song was liked by participants with varying moods.

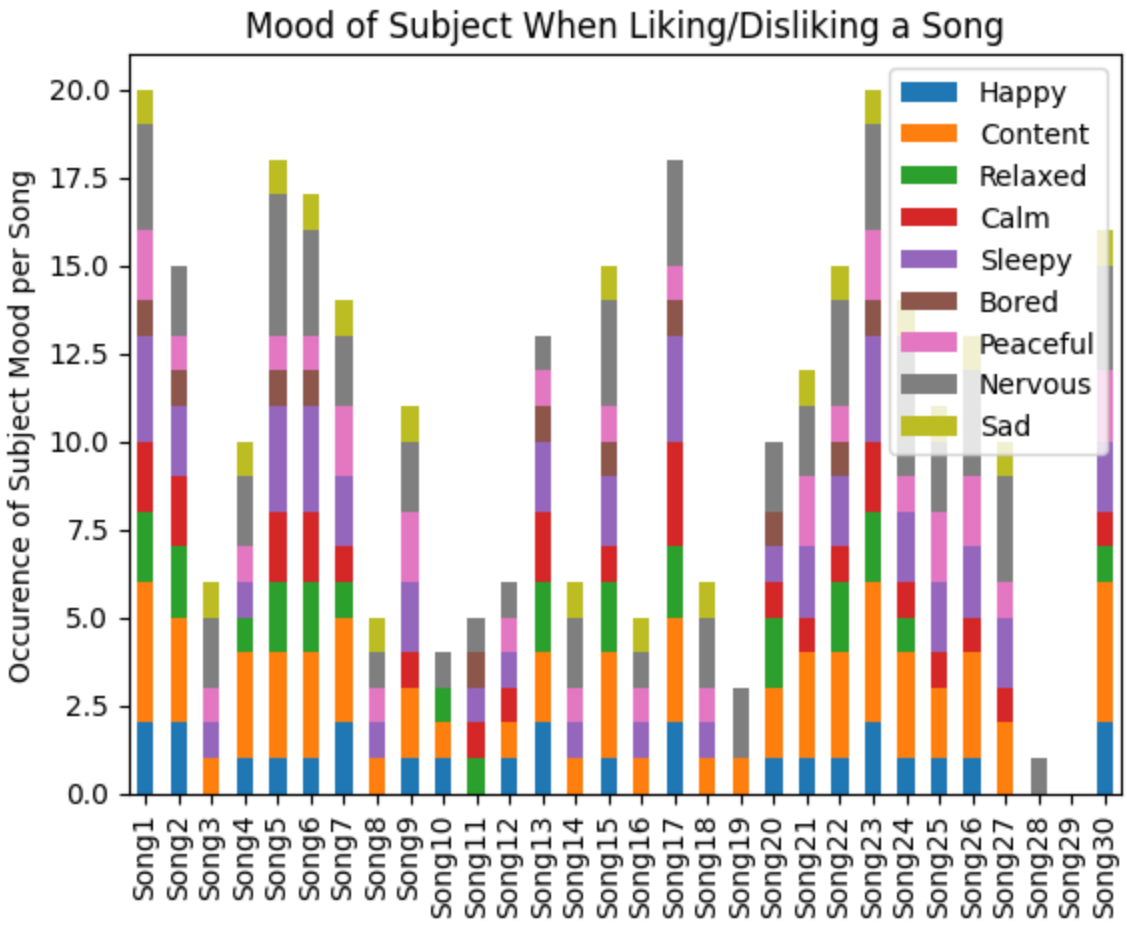

Figure 1. Mood of participant when listening to a song sample they liked

It was also important to make sure that all the participants did not like the same songs. The difference in song choice across subjects is presented in Figure 2. This figure shows the number of songs liked differently between each participant. The average number of differently liked songs between subjects was 13.33 with a standard deviation of 3.705. This difference appears large enough to represent different sets of music taste that might correspond with different sets of acoustic speech features.

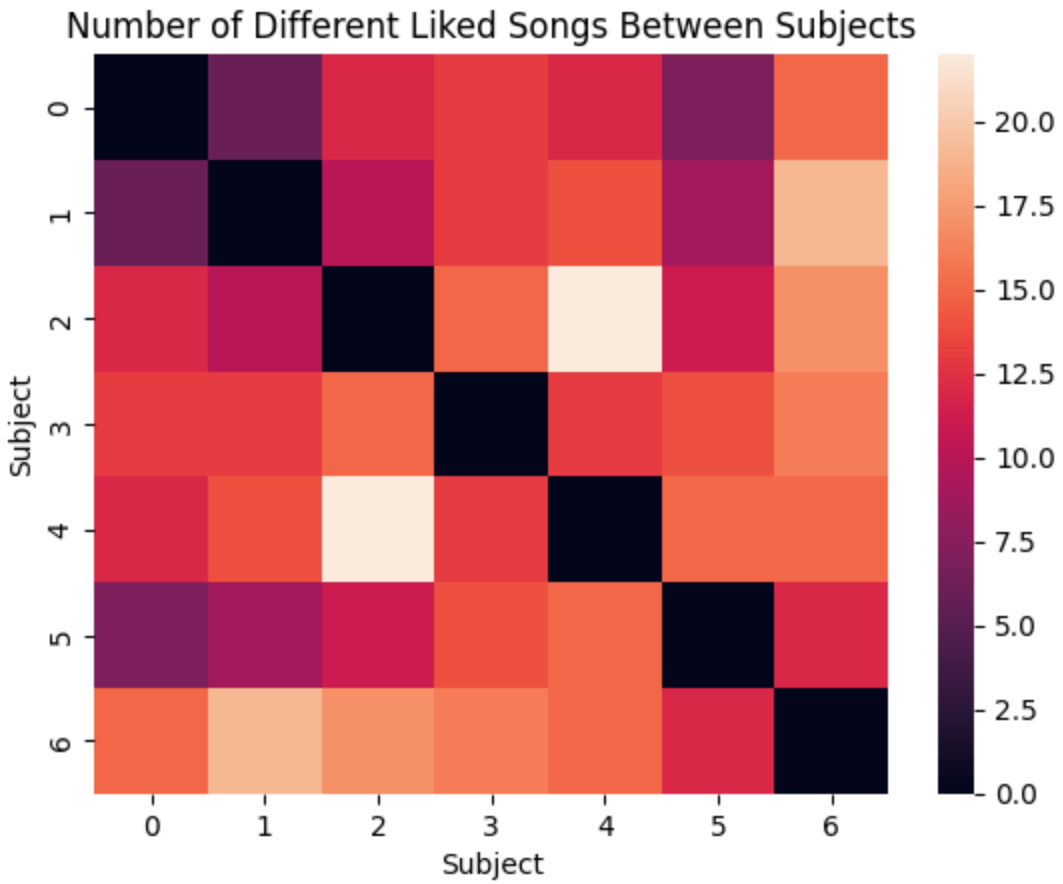

Figure 2. Number of songs liked differently between each participant

The absolute cross correlation matrix between features in participants speech and features in the songs they liked is displayed in Figure 3. This figure is a visualization of the cross correlation matrix detailed in 4.1 Analysis. The absolute value of the correlation coefficients was visualized rather than signed correlation coefficients because at this early stage of analysis, I am mainly concerned with finding any correlation between features in speech and music rather than the direction of the correlation. This way the image is less noisy and any structures in the correlations will be more evident. The rows in this image correspond with the correlation between a given feature in speech and features in music. The columns correspond to a given

music feature's correlation with features in speech. The largest magnitude cross correlation was found to be about .3.

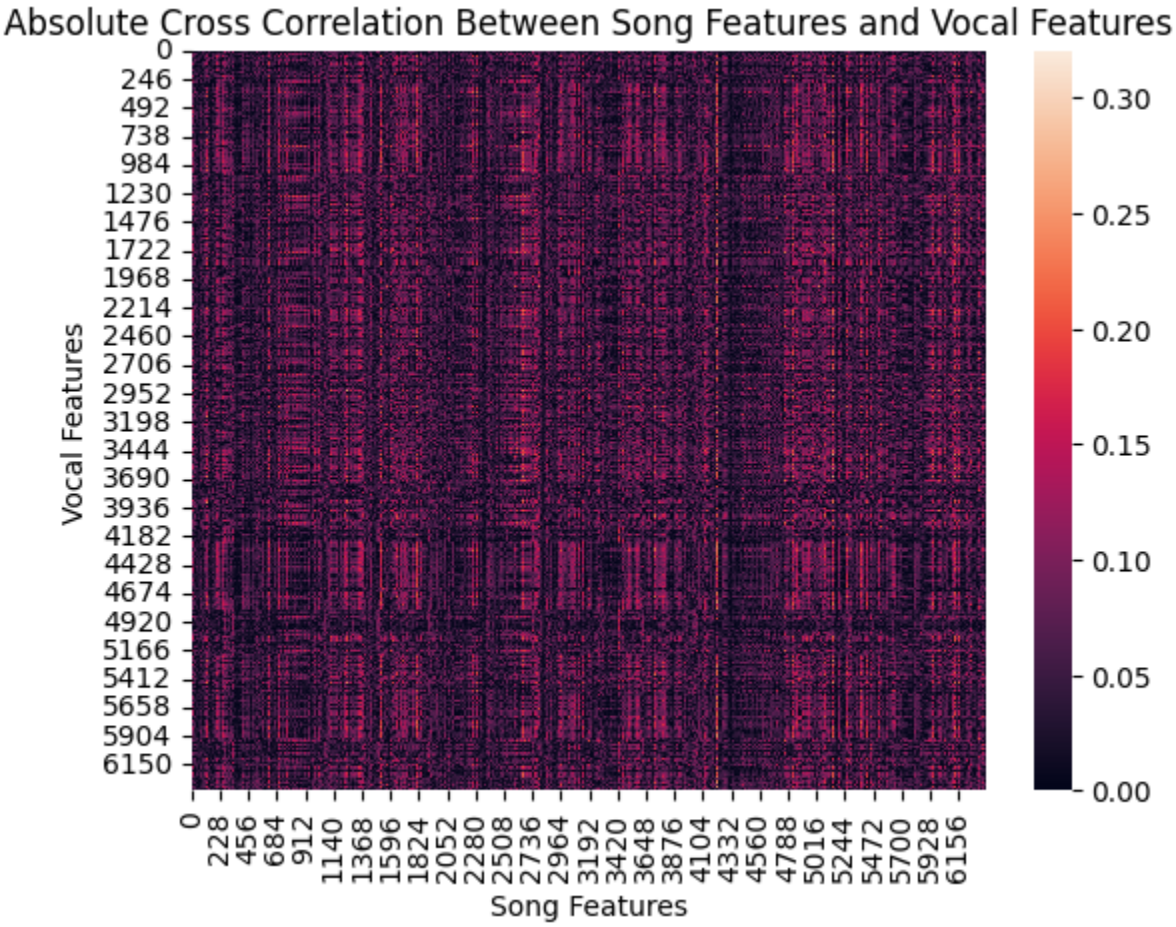

Figure 3. Absolute cross correlation between song features and vocal features

5: Discussion

The hypothesis that there are acoustic features in a person's speech that will correlate with the acoustic features in their preferred music was not supported by this study. As seen in Figure 3, the cross correlations between features in speech and features in music are all low. Figure 3 shows that the cross correlations can mostly be viewed as random noise. There does seem to be some vertical structure in the data with higher cross correlations found along bands of columns

such as between song features 4788 and 5700. However, given the lack of symmetry in this image, this is most likely due to a high variance for these song features rather than a correlation between the song features and speech features.

Although this study did not find any strong correlations between features in participants' speech and features in the songs they liked, it does not mean that a relationship does not exist. The low correlations can be due to a large number of factors. The most obvious being there were only seven subjects which makes any kind of statistical analysis very difficult. The choice in stimulus could have also affected the ability to find correlations, specifically the chosen songs. The lack of correlation could also be attributed to the fact that perception is complex and most likely can not be represented as a linear relationship between the features. The relationship between the features is presumably more complex. For this reason, it would be beneficial to try to get tens of thousands of participants and then use more advanced statistics to look for relationships.

This pilot study built on the trend of representing music and speech in the same feature space. This shared feature space was used to look for evidence of an influential relationship between speech perception and music perception. Although this study did not find any strong correlations that support this relationship, continuing this experiment and getting more participants could help to reveal more about the relationship. Ultimately, this pilot study helps to illuminate the ability and potential of technology to help us learn more about complex systems like music perception and perception in general.

References

Deutsch, D., Henthorn, T., & Dolson, M. (2004). Speech Patterns Heard Early in Life Influence Later Perception of the Tritone Paradox. *Music Perception*, *21*(3), 357–372. https://doi.org/10.1525/mp.2004.21.3.357

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on Multimedia*, 1459–1462. https://doi.org/10.1145/1873951.1874246

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*, 770–814. https://doi.org/10.1037/0033-2909.129.5.770

McDonald, G. (n.d.). *Every Noise at Once*. Retrieved December 15, 2022, from https://everynoise.com/engenremap.html

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*(3), 715–734. https://doi.org/10.1017/S0954579405050340

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., & Kim, S. (2013). *The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism*.

van der Schyff, D., & Schiavio, A. (2017). The Future of Musical Emotions. *Frontiers in Psychology*, *8*. https://www.frontiersin.org/articles/10.3389/fpsyg.2017.00988

Vuilleumier, P., & Trost, W. (2015). Music and emotions: From enchantment to entrainment. *Annals of the New York Academy of Sciences*, *1337*(1), 212–222. https://doi.org/10.1111/nyas.12676

Weinberger, S. (2015). *Speech Accent Archive. George Mason University.* http://accent.gmu.edu/

Weninger, F., Eyben, F., Schuller, B., Mortillaro, M., & Scherer, K. (2013). On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common. *Frontiers in Psychology*, *4*. https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00292

Wermke, K., Teiser, J., Yovsi, E., Kohlenberg, P. J., Wermke, P., Robb, M., Keller, H., & Lamm, B. (2016). Fundamental frequency variation within neonatal crying: Does ambient language matter? *Speech, Language and Hearing*, *19*(4), 211–217. https://doi.org/10.1080/2050571X.2016.1187903

Zhang, B., Mower Provost, E., Swedberg, R., & Essl, G. (2015). Predicting Emotion Perception Across Domains: A Study of Singing and Speaking. *Proceedings of the AAAI Conference on Artificial Intelligence*, *29*(1). https://doi.org/10.1609/aaai.v29i1.9334