Zain Souweidane

A Review of the Relationship Between Music and Language

Abstract

This review paper explores the relationship between language and the perception of music. It is conducted with the intention of highlighting papers that suggest that language influences the perception of music. This is done by reviewing studies that found similar acoustic features in music and language and shared neural processing of music and speech. Studies that observe correlations between musical abilities and speech processing abilities and the influence of language on the perception of music through cross-cultural and developmental studies are also reviewed.

Introduction

While our emotional response to music may be influenced by many experiences, this review will highlight studies that found an influential relationship between language and the perception of music. This review is not implying that language is solely responsible for our perception of music. It is reviewing the significant influence language and verbal communication has on shaping our emotional response to music. The influence of language on music perception helps support claims made about general perception across many disciplines. To keep the review simple, these claims are only briefly investigated in the background but require much more review to make broader claims.

To explore the influence of verbal communication on music perception, I first review the shared acoustic features of music and speech. Then I review studies that examine the overlap of the neurological processing of music and language. Next I review studies that observe a strong

correlation between music and language processing abilities. Finally, I review studies that suggest language influences perception of music.

Background/Motivation

Sounds heard in music are usually not directly heard in the natural world. Suppose the exact sound of every musical phrase could be found in everyday experiences. In that case, it could be hypothesized that music creates emotion because it reminds us of those experiences. However, the sounds heard and the emotions triggered by music are usually more general. Music evokes a wide range of complex emotions that do not always fall into basic emotional categories (van der Schyff & Schiavio, 2017; Vuilleumier & Trost, 2015). This complexity of 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 better understand our perception of music is the experience of language and vocal expression.

Theories on the relationship between music and language have been postulated in various disciplines, including psychology, neuroscience, evolution, and phenomenology. Many theories have attributed the similarities between music and language to coevolution, stating that they stem from affect bursts (Scherer, 2013) or represent the sounds of physiological processes, known as Spencer's Theory (Spencer, 1857).

The origin of music and the reason for similarities between music and language will likely not be known for quite some time. However, there is increasing evidence from studies confirming their correlation and suggesting that language and vocal expression can influence music perception.

This evidence agrees with phenomenological perspectives about experience shaping our perception of our senses, with this specific example highlighting a reason as to how language might influence the way we attribute meaning and emotion to sounds in music.

Shared Acoustic Features in Music and Language

If our emotional response to music is influenced by verbal communication, it would be logical to speculate that the reason for the relationship might lie in similar acoustic features between music and verbal communication. Furthermore, a set of acoustic features in verbal communicative phrases should evoke similar emotions to musical phrases with similar 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, 104 studies of vocal expression and 41 studies of music performance were reviewed. The review qualitatively related features in music to features 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, they found that studies that tested emotional response to music and studies that tested emotional responses to speech had similar results. 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 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 exposed acoustic features unique to music and speech.

One study selected a set of low-level acoustic features and created two nonlinear regression models using recurrent neural networks: one trained on musical samples to predict valence and arousal levels in musical samples and the other trained on speech samples to predict valence and arousal levels in speech samples (Coutinho & Dibben, 2013). The music model and speech model had five of the same input features: loudness, tempo/speech rate, melodic/prosodic contour, spectral centroid, and sharpness. Each model had one other distinct input feature: the music input vector included spectral flux, and the speech input included roughness. Table 2 displays the training and testing correlation of determination for the valence and arousal for both models.

	Valence	Arousal
Speech Training	.70	.89
Speech Testing	.63	.81
Music Training	.83	.97
Music Testing	.43	.75

Table 2. Training and Testing correlation of determination for the valence and arousal for the speech and music models from (Coutinho & Dibben, 2013)

These results show that reasonable models can predict emotion in music and speech using the same features: loudness, tempo/speech rate, melodic/prosodic contour, spectral centroid, and sharpness. This supports theories that emotional cues in speech and music are primarily encoded in the same acoustic features. Other interesting observations from this study are that valence is more consistently predictable in speech signals, and arousal is more predictable across both domains of music and speech.

An influential study demonstrated similar results using support vector machine models that 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 (Weninger et al., 2013). The effect of narrowing the ComParE Feature Set (H. Zhang, 2017) from 6373 features to 200 features yielded 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 these results, as well as the result that arousal is more predictable across both domains of music and speech (B. Zhang et al., 2015).

This theme of cross-domain analysis of models was continued in the subsequent years. Studies explored this idea by using various forms of transfer learning to compare the features in music and speech, with one study reproducing similar results when training and testing across domains (Coutinho & Schuller, 2017). Another showed improved accuracy of an emotion prediction model when trained on a large data set of speech and fine-tuned on a smaller music data set (Gómez Cañón et al., 2021).

Present technological capabilities validate the theory that music and speech share acoustic features which evoke similar emotions. This helps to support the idea that music and speech are deeply connected. Next, I will briefly review the neurological processing of music and emotion and how these features are processed.

Shared Neurological Processing of Music and Language

Before neuroimaging took off in the 1970s, the neural processing steps between stimulus and response were thought to be a localized structure that did not use a bilateral neural network. With the use of advanced technology, this notion has given way to the belief that in the presence of external stimuli, a distributed network of modules across both hemispheres with intrinsic properties integrate to form the steps between stimulus and response (Raichle, 2009; Sporns, 2013).

This breakthrough in multisensory processing has helped reformulate original ideas about speech and musical processing. Originally, speech processing had been thought to be controlled primarily by left-hemisphere auditory areas of the cerebral cortex (Wernicke, 1874). The ability to create and process music was attributed to the right side of the brain. This is due to the fact that the left hemisphere is relatively specialized for rapid temporal processing, whereas the right hemisphere focuses on spectral discrimination (Hickok & Poeppel, 2007; Zatorre et al., 2002).

Our original understanding of speech and musical processing was partially correct. A significant portion of speech processing occurs in the left hemisphere, and a significant portion of musical processing occurs in the right hemisphere. However, as suggested by the aforementioned breakthrough in multisensory processing, speech and music processing is less lateralized than once believed. Both have been found to use an expansive bilateral neural network, with the processing of musical features such as key and rhythm having been shown to engage several areas in both hemispheres (Alluri et al., 2012). This is further evidenced by a study that compared PET scans of subjects who generated melodies and sentences, revealing activations in nearly identical functional brain areas, including the primary motor cortex, supplementary motor area, Broca's area, anterior insula, primary and secondary auditory cortices, temporal pole, basal ganglia, ventral thalamus, and posterior cerebellum (Brown et al., 2006).

The shared use of functional brain areas in musical and speech processing shows that music and verbal communication share acoustic features. The overlap of neurological processing between music and speech also indicates the potential for both to have perceptual influences on each

other. The next section reviews studies that found correlations between musical abilities and language capacities.

Correlations between Language and Music Capacities

Numerous studies have shown that musical and speech-processing abilities are indicators of each other. Studies have shown this by observing speech processing abilities in musicians vs. non-musicians, music processing abilities in tone language speakers vs. non-tone language speakers, and language learning capacity of musicians vs. non-musicians.

Many studies have shown that increased musical abilities are associated with higher levels of speech processing. Specifically, they have shown that trained musicians were better at recognizing emotions in speech prosody compared to non-musicians (Lima & Castro, 2011). Another study found that musicians were more sensitive to a broader range of linguistic information, such as timing and spectral features (Sadakata & Sekiyama, 2011). As well as sensitivity to specific linguistic features, it has been proven that musicians can process speech in the presence of noise (Parbery-Clark et al., 2009).

Studies have also shown that the language a person speaks is related to their ability to process music. Tone languages are languages where pitch conveys lexical information, such as Cantonese and Mandarin. Pitch change in non-tone languages, such as English, does not convey lexical information but usually emotional meaning. Tone language speakers have been found to have an increased ability to process music. One study's findings which support this compared the ability to identify short melodies between non-musicians who spoke Mandarin and

non-musicians who spoke English (Alexander et al., 2008). This study found that the Mandarin speakers were better able to discriminate between melodies.

The ability to learn a language, both primary and secondary, has been shown to be related to musical processing abilities. Rhythm perception was found to be an indicator of grammar skills in developing children (Gordon et al., 2015). Musical skills are also related to the ease of learning the sound structure of a language when learning a second language (Slevc & Miyake, 2006)). This shows a correlation between musical ability and second language learning ability, but this might be a feature of shared neurological processing networks rather than a causal relationship (Zheng et al., 2022).

This section highlighted some of the many examples which indicate that musical abilities and perception are observed to be correlated with linguistic abilities and speech perception. The next section reviews studies that suggest language and vocal expression influence the perception of music.

Musical Perception is Developed and Reflects Language and Culture

The past sections provided evidence that language and music perception are strongly correlated and have the potential to influence each other. This section reviews studies that support theories that language and verbal communication have an influential relationship with music by observing that music perception varies across cultures. The studies reviewed highlight cross-cultural variation in emotion perception and expression, the development of music perception from an early age, and the influence of language on a culture's music.

Cross-Cultural Variation in Emotion Perception and Expression

Before looking at how the perception of music varies across cultures, studies that show the perception of emotions depends on cultural background were reviewed. How people express and perceive emotions vocally is not universal across cultures. This is due to many factors, including cultural norms, varying physiology, and the language of the culture. One study looked at how well people from Germany, Romania, and Indonesia were able to categorize emotional utterances in the German language (Jürgens et al., 2013). Recognition accuracy is higher when emotions are both expressed and perceived by members of the same cultural group—supporting the dialect theory, which states that an in-group advantage in emotion recognition exists. Jürgens et al. found that German subjects revealed a slight advantage in recognizing emotions. However, this study observed group emotional biases in the categorization of the utterances. German participants exhibited a further bias toward choosing anger for the utterances. Whereas Romanian and Indonesian participants were biased toward choosing sadness. These biases indicate that emotion perception is related to cultural background.

Cultural differences also affect vocal emotion perception in the absence of language (Waaramaa & Leisiö, 2013). This study conducted listening tests for 50 participants from the following five countries: Estonia, Finland, Russia, Sweden, and the United States. The participants were tasked with identifying emotions from speech samples containing only made-up words and sounds. Results showed that emotional identification largely depended on the participant's cultural background and age. The observation that emotional identification was dependent on age helps to support the idea that vocal emotional perception is flexible and developed.

More recent studies have continued to support the idea that vocal emotional perception is developed and influenced by language. Language and culture are important factors in developmental changes in the perception of facial and vocal affective information (Kawahara et al., 2021). This study tested participants' ability to determine whether vocal and facial expressions were congruent. The participants were East Asian (Japanese) adults, Western (Dutch) Adults, East Asian Children, and Western Children. Eastern and Western specifications are used to emphasize the cultural differences and follow studies that observed emotional perception in tone and non-tone languages. It was found that East Asian adults relied more on vocal cues than Western adults. However, young children from both cultural groups behaved like Western adults, relying primarily on visual information. Another relevant observation from the study was that the proportion of responses based on vocal cues increased with age in East Asian but not Western participants. This suggests that emotional perception and speech perception are influenced by cultural dialects and develop over time. These findings are consistent with findings that the development of music perception starts from an early age.

The Development of Music Perception from an Early Age

Since music perception is correlated to speech perception, as presented in the preceding sections, it would make sense that music perception is similarly shaped by language and developed over time. As confirmed by many studies, it has been demonstrated that music perception and speech perception start forming and are potentially influenced by language from a young age, sometimes as early as prenatally. It was famously shown that babies cry melodies shaped by the ambient spoken language they are exposed to while still in the womb (Wermke et al., 2016). More

directly related to music perception, infants have shown perceptual preferences for melodies and rhythms to which they were exposed prenatally (James et al., 2002).

A critical example of the influence of language on musical perception is the tritone paradox. It occurs when two tones related by a half-octave (or tritone) are presented in succession. The tones are constructed so their pitch classes (C, C♯, D, etc.) are clearly defined, but their octave placement is ambiguous. Some listeners hear the sequence of tones descending, while others hear it ascending. Previous studies have shown 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 how they perceive the pair of sequences, even if the listener is not fluent in this language (Deutsch et al., 2004). This highlights that language has the potential to influence music perception, and early life experiences critically influence the formation of music and speech perception.

Influence of Language on the Cultural Music

Some studies suggest that language influences a culture's music and individual music perception. It was found that the Scotch Snap, a musical rhythm characterized generally as a short beat followed by a long beat, was heavily prevalent in American and British music (Temperley & Temperley, 2011). It was hypothesized that this is because the English language consists of more short stressed syllables which reflect a similar pattern to the Scotch Snap. Another study found similar results when comparing common rhythms in Japanese and American music to the rhythmic features in Japanese and English (Sadakata et al., 2004).

The studies highlighted in this section establish that the cross-cultural variation in vocal emotional perception is also seen in cross-cultural perception of music. This influence on vocal emotional perception was expectedly reflected in musical perception on the individual level and the cultural level. These studies strongly suggest that music perception is developable, with language strongly influencing it. However, it is nearly impossible to confirm this influential relationship with absolute certainty due to the complexity of neurological processing and the complexity of language and music.

Conclusion

The phenomenological perspective views the world as a field for perception to which human consciousness assigns meaning. This review presented studies that suggest language and verbal communication influence how our consciousness assigns meaning to music. Studies have shown that music and speech contain comparable acoustic features which can evoke similar emotions. These features are similarly processed in the brain due to an overlapping neural network which could help to explain correlations between musical and language capabilities. Supported by the strong correlation between music processing and verbal communication, studies have investigated cross-cultural differences and the development of music perception to suggest that musical perception is influenced by language.

Due to the broad scope of this review, the proposition that language influences the perception of music has applications in many fields and disciplines. In the age of recommendation systems and getting more information from less, one exciting application is using language and vocal features to recommend music. As established in this review, there is evidence suggesting that music can

be thought of as a function of language and many other features. As evidenced by an active patent granted to Spotify titled “Identification of Taste Attributes from an Audio Signal,” which details the use of user voice recordings in their music recommendation algorithm, the idea that a person’s perception of music reflects aspects of their vocal features is already being commercialized (Hulaud, 2018).

There is mounting evidence suggesting that language can influence our perception of music and that this relationship might be bidirectional. This association between language and music has been studied for some time. However, as seen in the past 20 years, we will better be able to investigate this as our technological capabilities advance.

Bibliography

Alexander, J. A., Bradlow, A. R., Ashley, R. D., & Wong, P. C. M. (2008). Music melody perception in tone‐language‐ and nontone‐language speakers. *The Journal of the Acoustical Society of America*, *124*(4), 2495–2495. https://doi.org/10.1121/1.4782815

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., & Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage*, *59*(4), 3677–3689. https://doi.org/10.1016/j.neuroimage.2011.11.019

Brown, S., Martinez, M. J., & Parsons, L. M. (2006). Music and language side by side in the brain: A PET study of the generation of melodies and sentences. *European Journal of Neuroscience*, *23*(10), 2791–2803. https://doi.org/10.1111/j.1460-9568.2006.04785.x

Coutinho, E., & Dibben, N. (2013). Psychoacoustic cues to emotion in speech prosody and music. *Cognition and Emotion*, *27*(4), 658–684. https://doi.org/10.1080/02699931.2012.732559

Coutinho, E., & Schuller, B. (2017). Shared acoustic codes underlie emotional communication in music and speech—Evidence from deep transfer learning. *PLoS ONE*, *12*(6), e0179289. https://doi.org/10.1371/journal.pone.0179289

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

Gómez Cañón, J. S., Cano, E., Herrera, P., & Gómez, E. (2021). Transfer learning from speech to music: Towards language-sensitive emotion recognition models. *2020 28th European Signal Processing Conference (EUSIPCO)*, 136–140. https://doi.org/10.23919/Eusipco47968.2020.9287548

Gordon, R. L., Shivers, C. M., Wieland, E. A., Kotz, S. A., Yoder, P. J., & Devin McAuley, J. (2015). Musical rhythm discrimination explains individual differences in grammar skills in children. *Developmental Science*, *18*(4), 635–644. https://doi.org/10.1111/desc.12230

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), Article 5. https://doi.org/10.1038/nrn2113

Hulaud, S. (2018). *Identification of taste attributes from an audio signal* (United States Patent No. US9934785B1). https://patents.google.com/patent/US9934785B1/en

James, D. K., Spencer, C. J., & Stepsis, B. W. (2002). Fetal learning: A prospective randomized controlled study. *Ultrasound in Obstetrics & Gynecology*, *20*(5), 431–438. https://doi.org/10.1046/j.1469-0705.2002.00845.x

Jürgens, R., Drolet, M., Pirow, R., Scheiner, E., & Fischer, J. (2013). Encoding Conditions Affect Recognition of Vocally Expressed Emotions Across Cultures. *Frontiers in Psychology*, *4*. https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00111

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

Kawahara, M., Sauter, D. A., & Tanaka, A. (2021). Culture shapes emotion perception from faces and voices: Changes over development. *Cognition and Emotion*, *35*(6), 1175–1186. https://doi.org/10.1080/02699931.2021.1922361

Lima, C. F., & Castro, S. L. (2011). Speaking to the trained ear: Musical expertise enhances the recognition of emotions in speech prosody. *Emotion*, *11*, 1021–1031. https://doi.org/10.1037/a0024521

Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009). Musician Enhancement for Speech-In-Noise. *Ear and Hearing*, *30*(6), 653–661. https://doi.org/10.1097/AUD.0b013e3181b412e9

Raichle, M. E. (2009). A Paradigm Shift in Functional Brain Imaging. *Journal of Neuroscience*, *29*(41), 12729–12734. https://doi.org/10.1523/JNEUROSCI.4366-09.2009

Sadakata, M., Desain, P., Honing, H., Patel, A. D., & Iversen, J. R. (2004). A cross-cultural study of the rhythm in English and Japanese popular music. In Proceedings of the international symposium on musical acoustics (pp. 41-44). Nara, Japan: ISMA.

https://doi.org/10.1111/j.1467-9280.2006.01765.x

Sadakata, M., & Sekiyama, K. (2011). Enhanced perception of various linguistic features by musicians: A cross-linguistic study. *Acta Psychologica*, *138*(1), 1–10. https://doi.org/10.1016/j.actpsy.2011.03.007

Scherer, K. (2013). *Affect Bursts as Evolutionary Precursors of Speech and Music* (pp. 147–167). https://doi.org/10.1007/978-88-470-5424-0_10

Spencer, H. (1857). The origin and function of music. Essays, Scientific. Political, and Speculative, 2, 1857.

Slevc, L. R., & Miyake, A. (2006). Individual Differences in Second-Language Proficiency: Does Musical Ability Matter? *Psychological Science*, *17*(8), 675–681. https://doi.org/10.1111/j.1467-9280.2006.01765.x

Sporns, O. (2013). Structure and function of complex brain networks. *Dialogues in Clinical Neuroscience*, *15*(3), 247–262. https://doi.org/10.31887/DCNS.2013.15.3/osporns

Temperley, N., & Temperley, D. (2011). Music-Language Correlations and the "Scotch Snap." *Music Perception*, *29*(1), 51–63. https://doi.org/10.1525/mp.2011.29.1.51

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

Waaramaa, T., & Leisiö, T. (2013). Perception of emotionally loaded vocal expressions and its connection to responses to music. A cross-cultural investigation: Estonia, Finland, Sweden, Russia, and the USA. *Frontiers in Psychology*, *4*. https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00344

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

Wernicke, C. (1874). *Der aphasische symptomencomplex...* http://cesimadigital.pucsp.br/handle/bcd/2246

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, *6*(1), 37–46. https://doi.org/10.1016/S1364-6613(00)01816-7

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

Zhang, H. (2017, May 2). *2013 ComParE Feature Set*. HuisBlog. http://huisblog.cn/2017/05/02/IS13feature/index.html

Zheng, C., Saito, K., & Tierney, A. (2022). Successful second language pronunciation learning is linked to domain-general auditory processing rather than music aptitude. *Second Language Research*, *38*(3), 477–497. https://doi.org/10.1177/0267658320978493