The Perception of Emotion in Speech and Song

Camille Noufi
Center for Computer Research in Music and Acoustics, Stanford University

Our study aims to further understanding of emotion in speech and music by specifically analyzing perception of a hybrid of the two: the singing voice. Previous studies have focused on analyzing the accuracy and development of emotional perception in response to speech, music, and affective vocal bursts. Recently the RAVDESS dataset was released to further study of emotion in speech and singing. Our study aims aims to perform a simplified replication of the RAVDESS data collection study and integrate the analyses with the results of other speech, music and affect burst studies. We ask fourteen participants to classify spoken and sung phrases as having one of six emotions. Our results align with those found in the RAVDESS study and also support the results found by other vocal and musical burst studies that demographic information of both the speaker and the listener affect the perceptual interpretation of the vocalized emotion.

Introduction

The nuanced similarities between perception of speech and music have long been studied in the literature. A particular hybrid of both is singing. Is it more like speech, or more like music? Is it an even combination of both? Does it transcend the two to become an entity of its own? Clearly, singing defies simple categorization. Singing elicits a unique perceptual response as well. Song is one of the most historical and innate mediums of communication between humans. What can singing convey that speech alone does not quite capture? A mixture of music, semantics and acoustic prosody, the singing voice carries complex meaning through many vehicles, and therefore can be perceived in more than one way.

Previous research has aimed to understand the perception of emotion carried within speech and music. Some studies focused on understanding the potency of many different emotions portrayed through the same medium (e.g. speech) and try to define which emotions are the most easily understood in that medium. These studies typically create stimulus set of one sound type and vary the different emotions portrayed in this sound type. Listeners are asked to identify which emotion they perceive as present within the sound. Belin, Fillion-Bilodeau, and Gosselin (2008) sought to do this through the creation of the 'Montreal Affective Voices' (MAV) dataset and corresponding listening study. This study recruited ten actors (five male and five female) to each record nine non-verbal affective vocal bursts that corresponded to nine different emotions. The bursts ranged in length from 385 ms to 2,229 ms (correlating with emotion type). Thirty participants then listened to and rated the valence, arousal, intensity and emotion of each vocal burst via a forced-choice labeling of emotion and rating task of the other choices. The study found that emotions were able to be identified with a mean accuracy of 68%, with little confusion in identifying the negative emotions sadness, anger, and disgust. The most confusion happened between happiness and pleasure, fear and surprise. The study also examined the effects of actor and listener demographics, namely gender, on the ratings. Females actors produced bursts rated as greater intensity and arousal and smaller valence. Male listeners tended to rate emotions as having higher intensity. Most significantly, they found that women were most accurate in identifying the emotions of the affect bursts created by women, while men were least accurate in identifying bursts by other men.

A follow-up study by **?** used a similar design paradigm but instead focused on "musical emotional bursts" (MEB) as stimuli. The studies utilized 20 musicians (10 violinists and 10 clarinetist) to play a set of brief musical executions expressing the basic emotional states of happiness, sadness, fear and neutrality. These executions lasted for an average of 1.6 seconds each. Sixty participants were then asked to listen to and rate either valence or arousal or identify an emotion via forced-choice response. Forty vocal affect bursts (portraying only the four given emotions) from the previous study were also included and rated in the same manner. Vocal affect bursts were identified at a higher rate than musical bursts in this study, but the musical bursts were still identified very accurately, with a mean of 80.4%. The study found that timbre seemed to play a role in valence ratings. Vocal and violin stimuli received similar ratings corresponding to each emotion class (happy > neutral > fear > sad) and differed from the clarinet ratings (happy > fear > neutral > sad).

Others forms of study have sought to understand how one medium might carry an emotion differently from the other, mask it, confound it, or make it more easily perceived. One recent study looked at the similarities and differences in perceived emotion within speech and musical contexts over the

course of development. **?** examined the perception of emotion (via forced-choice response emotion identification) in speech, music and affect bursts among children and adolescents. The study found that accurate emotional perception refined with age, and accuracy of speech and music emotion perception increased at the same rate. Affect bursts were more accurately identified more quickly, supporting the idea that more primitive sounds with less complexity easier to identify. As one might expect, happiness and sadness were the easiest to identify at a younger age. Anger and fear were subsequently identifiable later in development.

Another recent study by **?** sought to create a stimulus set allowing for direct comparison between speech and song vocalizations. The study culminated in the release of the 'Ryerson Audio-Visual Database of Emotional Speech and Song' or RAVDESS. The stimuli consisted of match speech and singing segments of the same neutral North American English seven syllable sentences. Twenty-four professional actors were each asked to produce six different emotions of normal and strong intensity during both speech and singing production. The emotions were angry, fearful, sad, happy, calm and neutral. 319 participants listened to each vocalization and rated strength and genuineness along a numerical scale, as well as provided the perceived emotional label via forced-choice response. Across emotion classes, song and speech received similar accuracy ratings. Calm and neutral were portrayed in separate instances by the actors but rated as a single category by listeners.

The study by **?** provides a new medium follow-up on the MAV (Belin et al., 2008) and MEB (**?**) study results for the case of the speech-music hybrid, the singing voice. Our study aims to integrate the questions and hypotheses developed by the discussed previous research via analyzing perception accuracy of vocalized emotions in speech versus song. The study uses the RAVDESS dataset (**?**). Participants listened to speech and singing vocalizations and reported a sample's perceived emotion using a matched one-digit number(**?**). We investigate if there is a significant difference in the perception of negative emotions (low valence) versus positive emotions (high valence) when sung versus spoken as well as the effect of the perceived arousal level. The study also aims at analyzing the impact of listener-actor gender on the rating choices and its relation to the MAV and MEB demographic findings. Further, we include age and musical background in our demographic analysis. We predict that emotions traditionally categorized as having higher valence will be perceived more accurately in sung samples, while lower valence will be perceived more accurately in speech. In addition, we predict that females perceive female vocalizations significantly more accurately than men perceive the emotions of other men, musicians will have slightly higher perception accuracies, and that age will not be a significant factor in listener rating choices when all of the participants are adults.

Table 1
Stimulus parameters and their corresponding numeric label for use in analysis. Vocal channel, emotion and gender are balanced within dataset. Intensity level is randomly sampled from a uniform distribution.

Parameter	Numeric Label
Vocal Channel	1 = speech 2 = song
Emotion	1 = neutral 2 = calm 3 = happy 4 = sad 5 = angry 6 = fearful
Intensity	1 = normal 2 = strong
Actor Gender	1 = male 2 = female

Methods

Participants

The study recruited fourteen participants from the Music 251 course and CCRMA community. The participation criteria required normal hearing and to be above the age of 18.

Participant information was gathered via a demographic questionnaire given at the end of the listening experiment. Participants numerically entered their birth year, years of musical experience, first language (L1) and gender. The average age of a participant was 31 (SD = 9.49). The average years of musical experience was 17.86 years (SD = 8.45). A participant's L1 was classified either as 'North American English' (10 participants), 'Other English Dialect' (1 participant) or 'Non-English' (3 participants). Gender was classified either as 'male' (10 participants),'female' (4 participants) or 'prefer not to say' (0 participants).

Stimuli

The study used the audio-only subset of the complete RAVDESS set created by **?**. We refined this audio subset to use an equal number of speech and singing samples rated uttered by three female and three male actors. A mixture of normal and strong intensities were present in the subset. The angry, fearful, sad, happy, calm and neutral vocalizations were equally represented. Each stimulus audio file has a unique identifier.

The RAVDESS audio consists of mono-channel recordings, sampled at 48 kHz with 16 bits per sample. Each

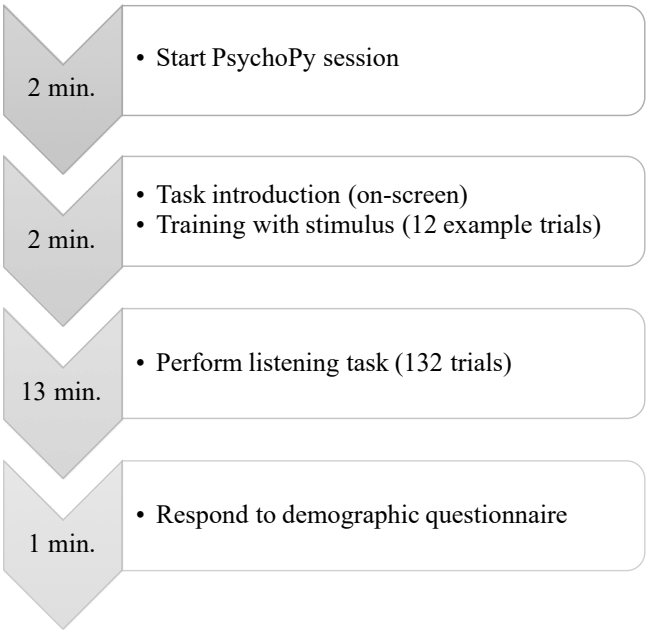


Figure 1. Diagram of listening task timeline for a single session. Each audio stimulus per trials lasts for approximately 5 seconds. Time estimates include response times.

stimulus (one per trial) is approximately 5 seconds in length. A full stimuli set presented in one session consisted of 132 trials, therefore a session length consisted of 11 minutes of audio. A description of the stimulus condition variations is presented in Table 1.

Procedure

A listening session took place within the CCRMA building at a 2013 MacBook laptop computer. The session program utilized a framework built in PsychoPy, an application for creating and executing behavioral science experiments in behavioral science (**?**).

A single evaluation session consisted of a participant listening to the complete set of vocalization samples described above in succession. The PsychoPy software was used to execute the session. Participants were given written instructions on how to start the evaluation task within PsychoPy. When beginning a session, listeners were provided with an on-screen description of their task. Then, six examples, one for each emotion and type of vocalization, were provided as training samples. Listeners were asked via on-screen prompt to categorize the sample's emotional category via labeled digits 1 through 6 via keyboard input. A 250 ms white Gaussian noise burst was played at -30dB before each trial. The training samples were not analyzed. The participant then received on-screen instructions that the full listening task would begin and subsequently listened to the remaining vocalizations, one vocalization per trial with a noise burst in between each trial. A session contained 132 trials. Each participant partook in a single session.

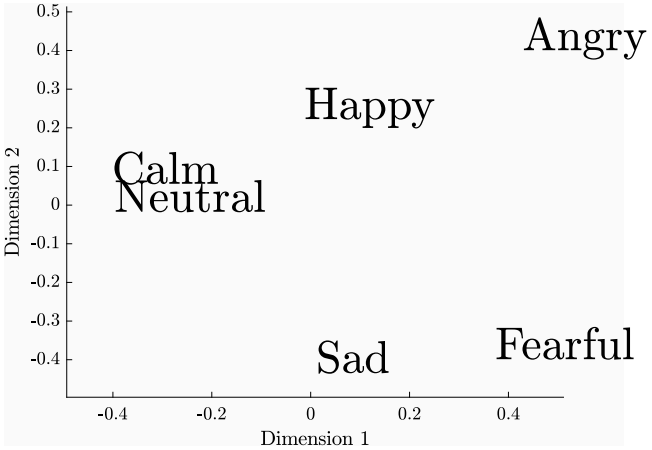


Figure 2. 2-D MDS plot of overall perceived emotion similarity by the participant group.

Following the listening session, the participants entered their demographic information. Entering this information concluded the evaluation task. Figure 1 presents the timeline of the task.

Data collection and Analysis

A listener's discrete rating responses were saved and stored for each trial. For a single trial, the numerical emotion label (1-6) was recorded alongside the unique stimulus identifier and existing metadata (see Table 1) of the played sample. The data is analyzed primarily at the group level, by categorical and demographic subgroups. At this level, confusion matrices and subsequent multidimensional scaling (MDS) plots are computed to analyze perception accuracy of different stimulus groups, and one-way analysis of variance (ANOVA) is used to investigate the effects of demographics on perception accuracy.

Results

Perception Accuracy by Stimulus Category

Confusion matrices are computed across all of the data and for four stimulus subgroups: speech, singing, normal intensity, and strong intensity. Hit-rate averages for each stimulus group are computed by averaging the normalized hit-rate accuracies of each emotion. A 2-D MDS plot (Figure **??**) is computed to capture the most prominent perceived similarity relations between the six emotions.

Figure **??** details the average perception hit-rate for each type of stimulus subgroup. The overall accuracy of emotion perception across all stimulus types is 60.5%. Accurate identification of emotion in speech is slightly higher, at 64.4%

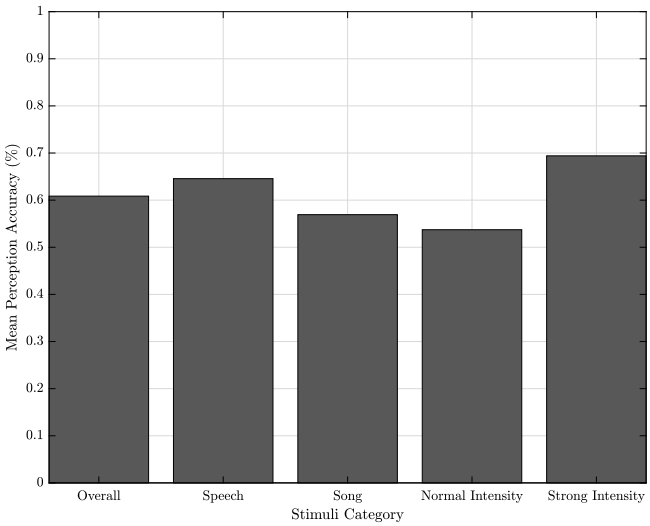


Figure 3. Group means of perception accuracy across all emotions, by stimulus category.

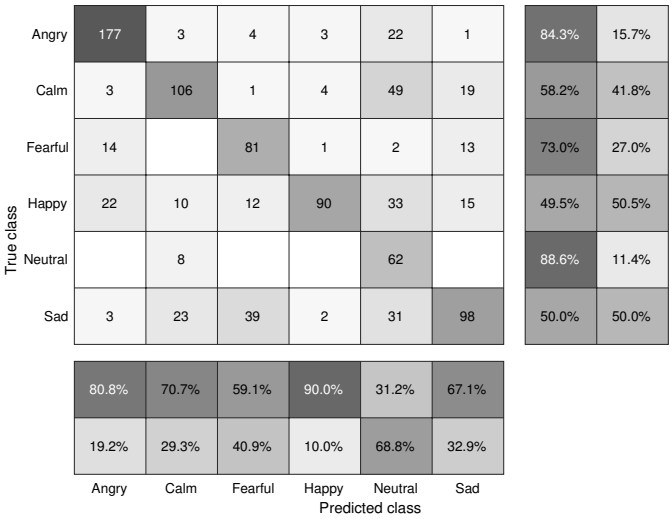
while identification of emotion in song is below, at 57.6%. Normal intensity emotions were the hardest to consistently predict, with a mean hit-rate of 53.1% while high intensity emotions were easier to identify at 69.6%.

Figure 4 shows the confusion matrices detailing the hit-rate accuracy and precision of emotion identification in speech and song. Neutral and angry are the easiest to identify in speech, with hit-rates of 88.6% and 84.3%, respectively. While happiness has the lowest hit-rate accuracy at just below 50%, it is the most precisely predicted, with a precision rate of 90.0%. In singing, anger is again the most accurately identified, at 80% correctness. Happy and sad follow in hit-rate accuracy, but with much lower rates (60.7% and 56.6%, respectively). Anger and fear are the most precisely identified in the context of singing, at 83.6% and 75.4%, respectively. Neutral emotion in singing is misclassified over half the time with even lower precision.

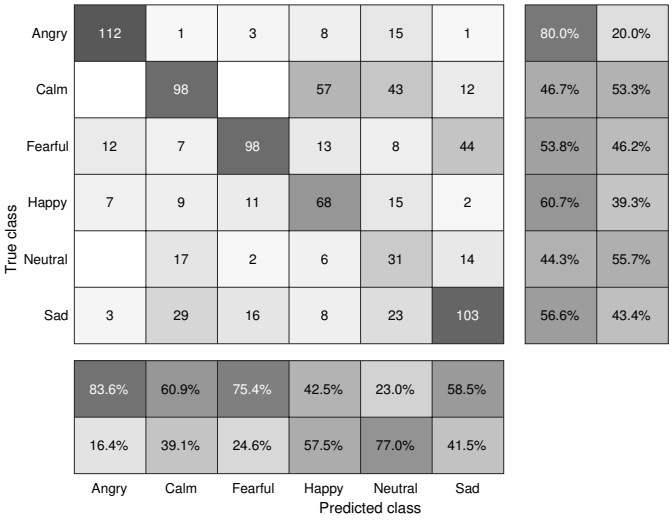
Perception Accuracy by Actor/Participant Group

Linear Pearson correlations of perception accuracy with birth year ($R = -0.197, p = 0.50$) and years of musical experience ($R = 0.32, p = 0.27$ appear to be insignificant for this participant cohort. Figure **??** shows each participant's overall perception accuracy in relation to birth year and self-provided years of musical experience.

One-way ANOVA is used to determine the influence of first language (L1) on vocalized emotion perception accuracy. Figure **??** shows the median accuracy of overall perception as a function of L1. Although there is a trend indicating a decrease in perception accuracy as the first language becomes more distant from North American English, the trend is not found to be significant ($F = 0.33, p = 0.73$).



(a) Speech.



(b) Singing.

Figure 4. Confusion matrices of group perception accuracy of vocalized emotions. Center confusion matrix represents count of accurate perceptions. Right column represents hit accuracy (%). Bottom rows represent precision accuracy (%).

Effects of gender were analyzed via computing the mean hit-rate accuracies of both listener and actor gender subgroups and via one-way ANOVA of accuracy as influenced by listener gender. Both analysis methods show slightly higher accuracies when either the speaker or listener is female. However, the ANOVA test shows the difference is not significant within this study ($F = 1.90, p = 0.19$). Figure **??** presents these summary statistics.

MDS plots are also computed for both listener and speaker gender subgroups. Figure 8 visualizes the perceived emotional similarity based on these four subgroups. The similar-

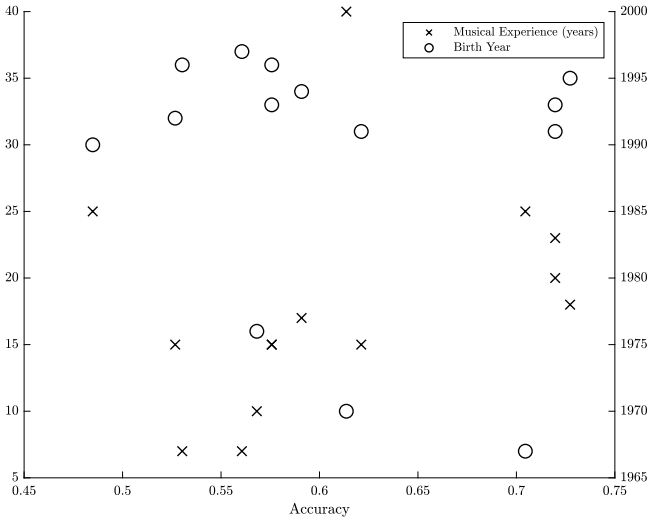


Figure 5. Scatter plot of overall emotion perception accuracy by a single listener as in relation to his/her birth year and years of musical experience. One '*O*' or '*x*' represents one listener.

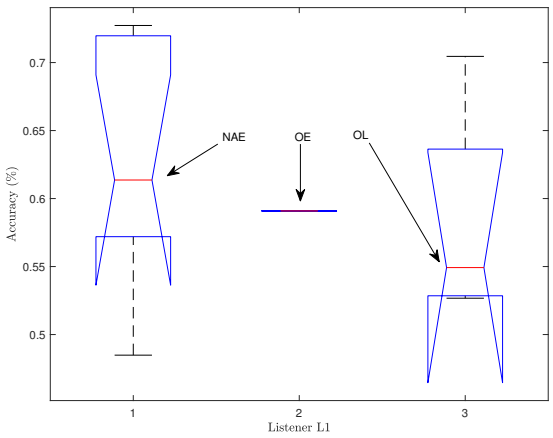


Figure 6. Group-wide median and inner-quartile range of perception accuracy as a function of first language (L1). 'NAE' = North American English, 'OE' = Other English Dialect, 'OL' = Other first language. Red horizontal lines represent each group's median. Horizontal box edges represent the 1st and 3rd quartiles. Whiskers extend to group outliers.

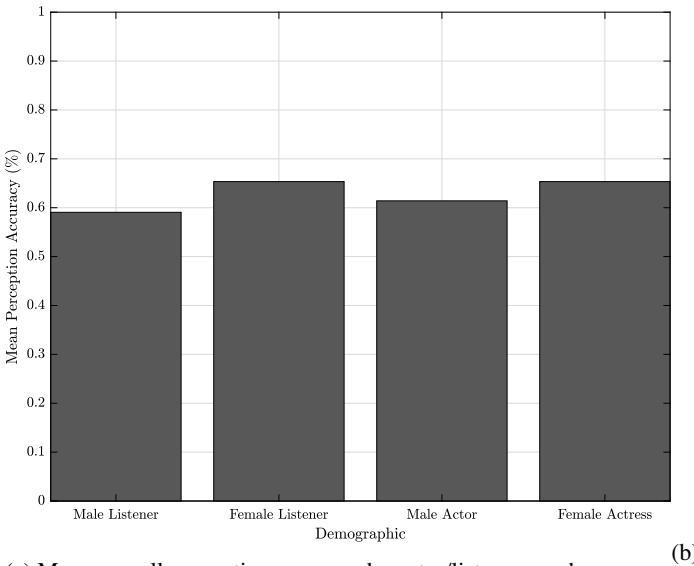
ity mapping for female listeners and speakers are very similar to the overall similarity (see Figure **??**. Similarities perceived by male listeners are similar but the second dimension appears to be flipped. Emotions portrayed by male actors are the most dissimilar from the overall perception, with happy and sad as well as neutral and calm seeing little separation across the two most prominent axes.

Discussion

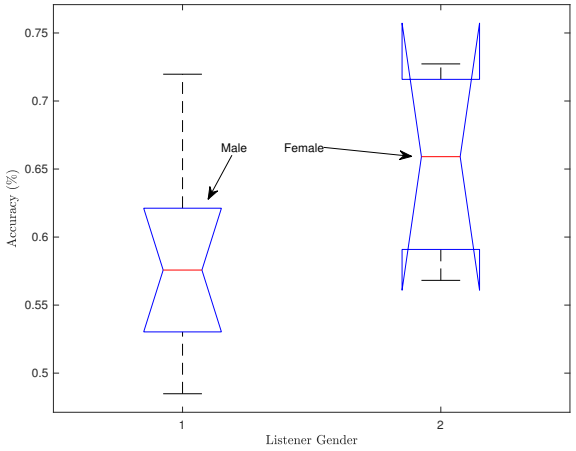
The results obtained in this study aligned with results from much of the previous literature. Emotions in speech were correctly identified more often than in song, and normal intensity emotions were much harder to identify than high intensity emotions. The easiest emotion to identify was anger, an emotion typically classified as high-arousal low-valence. Age and musical experience were not found to be influencing demographic factors for our participant group. The slightly higher accuracy results shown by female vocalizers and listeners as well as native North American English speakers were not found to be significant among this participant cohort.

Anger, a high-arousal emotion, makes sense as being the most accurately identified emotion. This result aligns with the fact that all emotions were easier to identify at a stronger intensity. In addition, the most precisely identified emotion was happiness, an emotion of both high arousal and high valence. The hit-rate accuracies for both singing and speech ordered similarly to those found by **?**. For speech, the accuracy rankings were almost identical, with neutral being the most identified, followed by anger, then fear, then calm, then happy and sad. Participants in our study identified neutrality in the singing voice more accurately than in the study by **?**. However, the perception accuracies of other emotions were ranked similarly: anger then happiness, sadness, fear and finally calm. Intensity of emotion also played a similar role in our study as in the study by **?**. The full study found a 5% improvement in perception accuracy when rating the high intensity emotions. Our study found an even more drastic difference between perception of normal and high intensity vocalizations, at a 16.5% improvement in accuracy. This could be due to the smaller sample size of both participant ratings and number of trials heard per session. Overall, high-arousal seems to be the most salient dimension of emotional classification in speech. Perception of valence in singing is more easily identified than in speech, likely due to valence information portrayed by presence of melody.

Our results differed greatly by emotional ranking from the results found by Belin et al. (2008) and **?**. Their results found that happy and sad were the easiest emotions to identify from vocal bursts, angry and fearful less so. The ranking is almost reversed. This difference in results suggests that vocal context length may play a role in ease of identification. Although our results were not statistically significant, we did see a similar similar 6% disparity between female and male listener hit rates. Likewise, the slight increase in accuracy of emotions as portrayed by female actresses over male actors suggests that the strongest accuracies would be by female listeners perceiving vocalizations by female actresses,as found in the

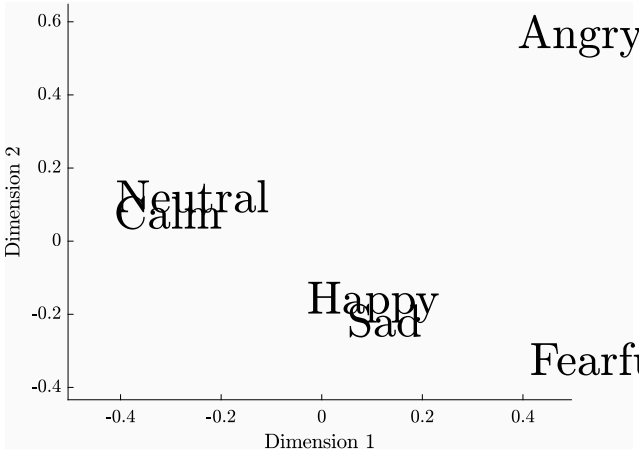


(a) Mean overall perception accuracy by actor/listener gender.

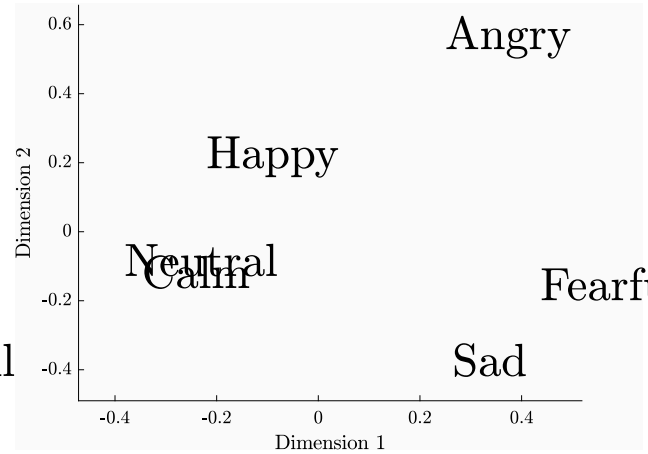


(b) Median perception accuracy by listener gender.

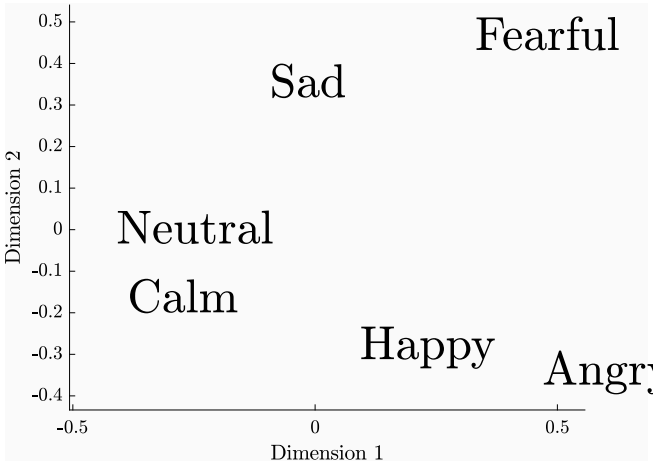
Figure 7. Group perception accuracy by gender.



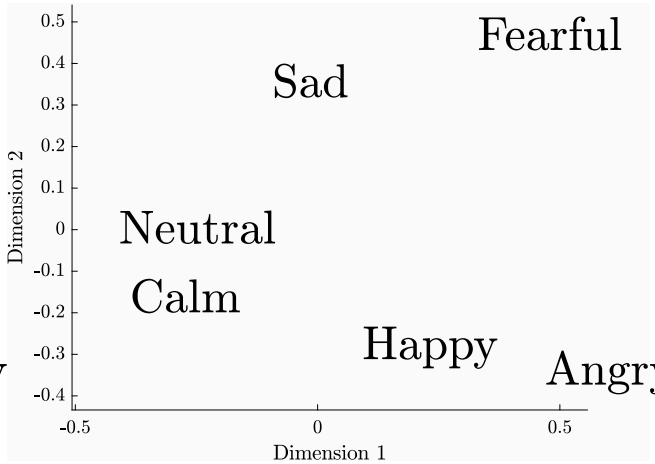
(a) Portrayed by male actor.



(b) Rated by male listener.



(c) Portrayed by female actor.



(d) Rated by female listener.

Figure 8. Perceived similarity of vocalized emotions, by gender.

MAV study (Belin et al., 2008).

The slight (but not significant) advantage North American English speakers appear to have in perception accuracy might be due not to language itself but cultural identification. All of the actors are native North Americans with North American English and their first language. Their representations of emotions are likely tied to their culture, as well as their association of melody to emotion in the singing passages. Finally, our results showed no significant relationship between birth year and perception accuracy amongst an adult population. This aligns with the results by **?** suggesting that emotional perception refines with age up until late adolescence, at which point perception accuracy reaches full maturity. While musical experience did not have a significant effect on perception accuracy, this could be due to the small participant numbers and therefore the large effect of outliers. Removing outliers, a visual trend is present in the data (see Figure **??**), and the correlation coefficient *R* is moderate, at 0.32. A larger participant cohort would allow for further study of the significance of this demographic effect.

Although many of our initial results do align with results from the discussed literature, it should be noted that our study used a much smaller group of participant than either the MAV, MEB or RAVDESS data collection studies. Therefore, influences of outliers more greatly affect group means and variance in the data and significance values cannot be directly compared. Future work to collect more sessions of this data from many more participants would greatly improve the analyses and refine the initial conclusions drawn within our study, especially subgroup trends that were not found to be statistically significant.

Because many of our results align with previous perceptual work, an interesting extension of this type of research would be connecting it to automatic classification of vocalized emotion. Automatic vocal emotion recognition is an active research field, with state of the art models still performing with modest or poor classification accuracies. A study by **?** investigated the role of acoustic parameters in predicting and classifying ten emotions and neutral voice during singing. The study used phrases sung by professional opera singers. Results showed that valence and arousal were key dimensions implicit detected by the classifier to determine the sung emotion. It would be interesting to run an acoustically trained classifier over this dataset and compare the confusion matrices of predictions determined by the automatic classifier to the predictions done by humans. It is clear from our study and the ones previously discussed that emotional perception is not clear cut, and a more important question might be to ask what are the salient features of the voice that lead to a perceived emotional tone, and if these salient vocal features are similar or different depending on who (or what) is on the perceiving end of the emotional exchange.

Conclusion

Our study analyzed the accuracies of perceptual emotion recognition in response to song and speech. Our results aligned with the emotion recognition accuracies found in the much larger scale RAVDESS study by **?** and we extended this study to also support findings by Belin et al. (2008) and **?** that speaker and listener demographics influence the perception of vocalized emotions. The study suggests that strong emotional arousal is more influential than emotional valence in accurate perception and that valence may be easier to identify in song than speech due to the added cues provided by melody. We hope to expand our study to include a larger number of participants in order to validate our initial findings and understand the strength of their alignment with the existing literature.

References

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods*, *40*(2), 531–539. doi: 10.3758/BRM.40.2.531

Eyben, F., Salomão, G., Sundberg, J., Scherer, K., & Schuller, B. (2015). Emotion in the singing voice—a deeperlook at acoustic features in the light ofautomatic classification. *Eurasip Journal on Audio, Speech, and Music Processing*, *2015*(1). doi: 10.1186/s13636-015-0057-6

Livingstone, S. R., & Russo, F. A. (2018a). *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).* doi: 10.5281/zenodo.1188976

Livingstone, S. R., & Russo, F. A. (2018b). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE*, *13*(5), 1-35. doi: 10.1371/journal.pone.0196391

Paquette, S., Peretz, I., & Belin, P. (2013). The 'musical emotional bursts': a validated set of musical affect bursts to investigate auditory affective processing. *Frontiers In Psychology*, *4*(509). doi: 10.3389/fpsyg.2013.00509

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. doi: 10.3758/s13428-018-01193-y

Vidas, D., Dingle, G. A., & Nelson, N. L. (2018). Children's recognition of emotion in music and speech. *Music & Science*, *1*. doi: 10.1177/2059204318762650