Temporal Response Functions Show that Collective Neural Processing Differs along the Speech-Music Continuum

Emily Graber[1], Malcolm Slaney[1,2], Shihab Shamma[3]

[1]Center for Computer Research in Music and Acoustics, Department of Music, Stanford University
[2]AI Perception, Google
[3]Department of Electrical and Computer Engineering and Institute for Systems Research, University of Maryland

Poster Summary

Decoding a stimulus from single-trial EEG is a topic of ongoing research
- We aim to understand how music and speech are decoded
- We explore how semantic category and acoustics affect the decoders and the decoding accuracy

We hypothesize that decoding models from within a semantic category will be similar, while decoding models in different semantic categories will be dissimilar. Further, we hypothesize that envelope prediction accuracy will be higher within category than across category.

Background

Temporal Response Functions

- A backward temporal response function (TRF) is a model mapping a neural response to an acoustic feature. It can predict stimulus features from neural data.
- Linear, envelope-based TRFs are outperformed by models using more complex features (band-limited envelopes, phonetic features) [1], however the improvements are small.
- Most studies using EEG focus on decoding speech [2-5], however music is a ubiquitous signal which has received less attention, but features prominently in daily life.

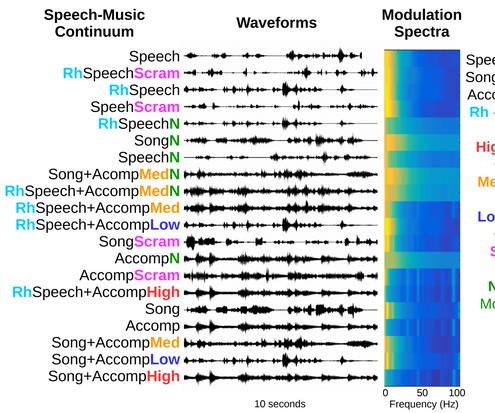
Music versus Speech

- Song and speech are semantically different, but they can have acoustic similarities as speech-to-song stimuli do. Still, during song perception, neural processing [6,7] and perception of pitch and time [8] may be different from that during speech perception.
- Instrumental music generally has temporal modulations reliably different from speech [9], as well as finer pitch differences [10,11] which may engage parts of the brain differently from speech [12].
- A capella song and instrumental music are categorically similar, yet acoustically different.

Speech-Music Continuum

We developed a set of stimuli ranging from speech to music. Some stimuli share acoustic features but differ semantically while the reverse is true for others. Time-scrambled controls preserve spectral content, but destroy temporal organization. Modulated pink noise controls remove spectral differences but preserve temporal structure.

Stimuli



contact: emgraber@ccrma.stanford.edu

Materials & Methods

Stimuli

Speech-Music Continuum
- A trained singer and pianist were separately recorded on the same stage with two Sennheiser cardioid condenser microphones. The pianist played the accompaniment part for two songs. The singer spoke, rhythmically spoke, and sang the lyrics for the same two songs.
  - For each song, their recordings were combined in 10 ways
    - relative amplitude envelopes manipulated by 12-15 dB
  - 6 pink noise modulated controls were computed for each song using only unique amplitude envelopes
    - envelopes were computed using RMS in 30 ms windows with 10 ms hop size
  - 4 additional controls were created from the solo recordings by time scrambling segments 50-200 ms long
    - scrambled segments taken from both songs
    - segment lengths drawn iid from uniform distribution

EEG

Participants
- 5 subjects (1 female), mean age 28.2 (SD = 1.8), mean years musical training 16.6 (SD = 6.5), NH

Task
- listen to and rate each clip on a scale from 1-7, where 1=Speech and 7=Music. This scale was used for all stimuli including controls

Recording
- 64 channel, Quik-Cap Neuroscan System with Curry7 acquisition software
- 500 Hz sampling rate with low pass at 200 Hz on-line, average reference off-line
- preprocessing done with Brainstorm in MATLAB (bandpass 1-32, downsample 64, model+remove blink SSP components, epoch)

Modeling

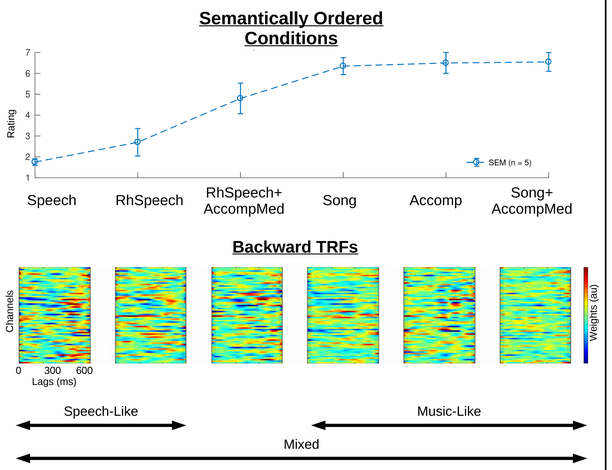
Backward TRFs (mapping EEG to audio envelope) [13]
- linear
- shrinkage regularization

Training
- Across two songs, all audio stimuli pairs concatenated (only scrambled stimuli not paired) resulting in 20 4-minute clips
  - RMS Amplitude envelope extracted for each clip in windows of 47 ms with 15 ms hop size. This is equivalent to 64 Hz sampling rate, z-scored
- Corresponding EEG data low passed to 9 Hz, and lagged to 600 ms in 15.6-ms steps
  - 10th order butterworth, filtered forward and backward

Evaluation
- TRF channel-wise similarity
  - within speech-like and music-like categories
  - assessed channel by channel with Pearson correlation
- Prediction accuracy
  - Assessed by each model's ability to predict all stimuli from real EEG
  - Pearson correlation used between predicted and actual amplitude envelopes

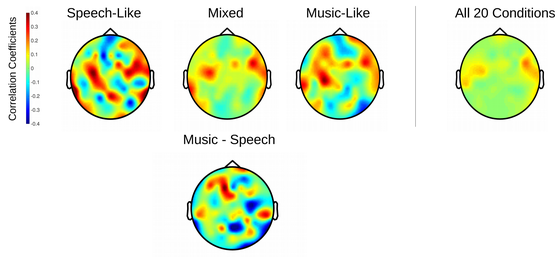
Results

Human judgments define a speech-music continuum



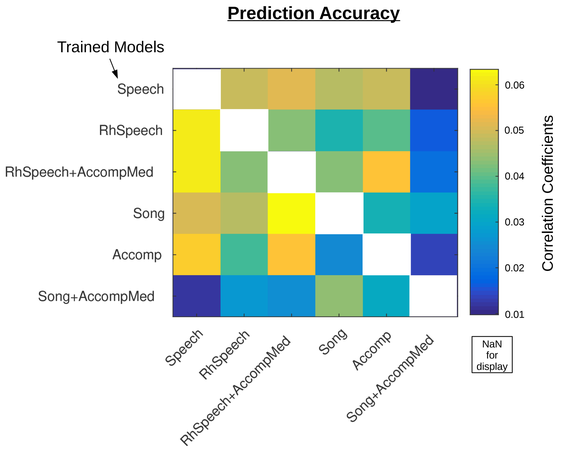
Results Cont'd

TRF channel-wise similarities in both the speech-like and music-like categories reveal consistent channel weights around C3 and FC6. Compared to speech-like models, music-like models are more consistent around F1, FCz, FC1, TP8, TP7.



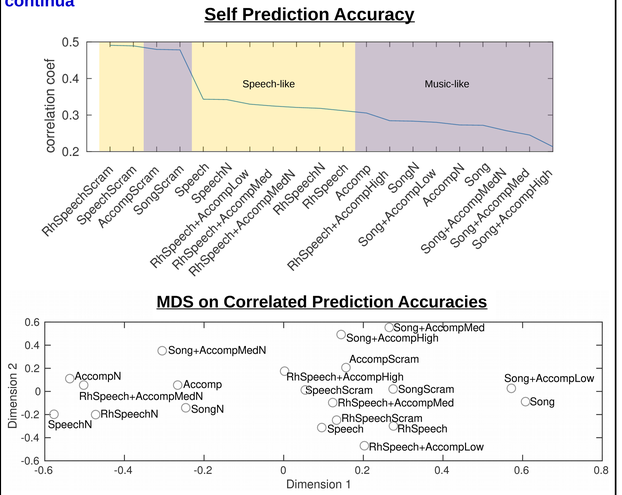
Envelope prediction accuracy, using each trained model to predict each condition, demonstrates that speech models are more general than music models. Additionally, the asymmetries may indicate how fine tuned each model is to it's own condition despite some generality.

- Ordering conditions acoustically (peak of modulation spectrum (MS), MS centroid, envelope MS centroid) produces less clear patterns.



Results Cont'd

Self prediction accuracy and MDS may define different speech-music continua



Discussion

- TRFs for decoding speech may not be good for decoding music; it might depend on how many parts are in the music
- Despite the hierarchical temporal structure of music, the music models are not good at decoding any stimuli
- The best ordering may be based on semantics rather than acoustics. The modulation spectrum peak for each condition does not correlate with the behavioral ratings
  - Some of the stimuli here could have been outliers.
  - The centroid works better, so peak value is not necessarily valid
- Future work can determine whether it is possible to classify speech or music using discriminative channels
  - overall best channel (C3, FC6) are the same for speech-like and music-like models
  - music training level was classifiable from fMRI with higher-level musical features [14]
- Control stimuli can be investigated further in the future.
  - Modulated noise models may not decode a real stimulus envelope well because of differences in the underlying neural processes
  - Scrambled conditions decoded themselves well, but unclear how they generalize
  - models can be averaged into different groups to determine more relevant 'dimensions' underlying the models

References

[1] Di Liberto, O'Sullivan & Lalor, (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. Current Biology, 25(19), 2457–2465.
[2] Aiken & Picton, (2008). Human cortical responses to the speech envelope. Ear and Hearing, 29(2), 139–157.
[3] O'Sullivan, Power, Mesgarani, Rajaram, Foxe, Shinn-Cunningham, Slaney, Shamma & Lalor, (2014). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. Cerebral Cortex, 25(7), 1697–1706.
[4] Mesgarani & Chang, (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature, 485(7397), 233–236.
[5] Akram, Presacco, Simon, Shamma & Babadi, (2015). Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. NeuroImage, 124, 906–917.
[6] Merrill, Sammler, Bangert, Goldhahn, Lohmann, Turner & Friederici, (2012). Perception of words and pitch patterns in song and speech. Frontiers in Psychology, 3(Article 76), 1–13.
[7] Tierney, Dick, Deutsch & Sereno, (2013). Speech versus Song: Multiple Pitch-Sensitive Areas Revealed by a Naturally Occurring Musical Illusion. Cerebral Cortex, 23, 249–254.
[8] Graber, Simchy-Gross & Margulis, (2017). Musical and Linguistic Listening Modes in the Speech-to-song Illusion Bias Timing Perception and Absolute Pitch Memory. The Journal of the Acoustical Society of America, 142(6), 3593-3602.
[9] Ding, Patel, Chen, Butler, Luo & Poeppel, (2017). Temporal modulations in speech and music. Neuroscience and Biobehavioral Reviews.
[10] Tillmann, (2014). Pitch processing in music and speech. Acoustics Australia, 42(2), 124–130.
[11] Zatorre & Baum, (2012). Musical melody and speech intonation: Singing a different tune? PLoS Biology, 10(7), e1001372.
[12] Abrams, Bhatara, Ryali, Balaban, Levitin & Menon, (2011). Decoding temporal structure in music and speech relies on shared brain resources but elicits different fine-scale spatial patterns. Cerebral Cortex, 21(7), 1507–1518.
[13] Akram, de Cheveigne, Diehl, Hjortjaer, Mesgarani, Parra, Shamma, Slaney & Wong, (2015). Telluride Decoding Toolbox. Institute for Neuromorphic Engineering, http://www.ine-web.org/software/decoding.
[14] Saari, Burunat, Brattico & Toiviainen, (2018). Decoding Musical Training from Dynamic Processing of Musical Features in the Brain. Scientific Reports, 8, 708(1-12).

This poster was presented at the 2018 Mid-Winter Meeting of the Association for Research in Otolaryngology, San Diego