Investigating the cortical tracking of speech and music with sung speech

Giorgia Cantisani, Amirhossein Chalehchaleh, Giovanni Di Liberto, Shihab Shamma

Investigating the cortical tracking of speech and music with sung speech

Giorgia Cantisani[1], *Amirhossein Chalehchaleh*[2], *Giovanni Di Liberto*[2,*], *Shihab Shamma*[1,3,*]

[1]Laboratoire des systémes perceptifs, DEC, ENS, PSL University, CNRS, France
[2]ADAPT Centre, School of Computer Science and Statistics, Trinity College, The University of Dublin; Trinity College Institute of Neuroscience, Ireland
[3]University of Maryland College Park, USA
[*] *senior authors*

giorgia.cantisani@ens.psl.eu

Abstract

The cortical tracking of speech and music has been primarily investigated separately. Here, we propose a novel paradigm involving sung speech to systematically compare the cortical encoding of sung speech with that of speech and music alone, offering a benchmark for using it in auditory research with ecologically-valid tasks. While this approach will ultimately lead to a variety of neural indices of speech and music processing at various levels of abstraction, the first step is to examine the envelope tracking of sung speech. EEG is recorded from subjects listening to a set of stimuli explicitly designed and built for the comparison: hummed melodies, speech monologues, and sung speech sharing the lyrics with the speech condition and the melody with the music condition. Preliminary analyses using encoding and decoding modeling show robust and consistent acoustic responses across conditions, with the only significant differences exclusively due to melody processing.
Index Terms: speech&music, auditory neuroscience, EEG

1. Introduction

Recent advances in auditory neuroscience, particularly in spoken language, led to a rapid increase in the use of linear modeling techniques for studying the tracking of natural stimuli in cortical signals such as electroencephalography (EEG) [1]. Unlike event-related potentials, which measure the average neural response to a discrete event, linear models seek to capture how changes in a particular stimulus dimension are linearly reflected in the brain activity. In other words, brain responses can be modeled as a linear combination of selected stimulus features, enabling the neurophysiological interpretation of the model weights and insights into the neural encoding of specific stimulus dimensions [2]. For instance, it has been consistently shown that cortical signals robustly track energy fluctuations of auditory inputs, referred to as amplitude envelopes. Such tracking is defined as cortical entrainment in the broad sense [3] and has been used as a measure of speech cognition (*e.g.*, [4]) and to explore new brain-computer-interfaces (BCIs) [5].

Cortical tracking can be measured using regularized regression, which can be intended as either a forward encoding or a backward decoding [6] (Figure 1). Forward encoding models predict the neural response at each channel as a weighted sum of time-lagged features of the auditory signal, giving physiologically interpretable traces called Temporal Response Functions (TRFs). This modeling of neural responses allows locating channels and stimulus-response latencies where the information of interest (*e.g.*, sound envelope) is encoded. The predicted neural activity is then correlated with the measured one to obtain a summary measure of cortical tracking. On the other side, backward models aim at reconstructing features of the auditory input

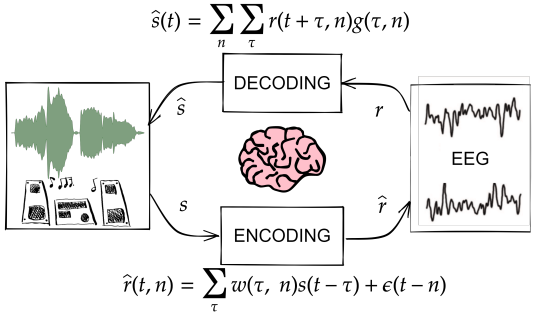

Figure 1: *The stimulus-response mapping function can be modeled in the forward direction (i.e., encoding) or backward direction (i.e., decoding), offering complementary ways to investigate how stimulus features are encoded in neural responses.*

as a weighted sum of the time-lagged neural response.

This framework can be used to investigate typical neural processing of speech [4], as well as cognitive and sensory deficits (*e.g.*, dyslexia, autism spectrum disorder, hearing impairment) [8]. While previous studies largely focused on speech listening tasks, recent applied work highlighted the need for more engaging paradigms, such as sung speech, when considering particular cohorts, from typically developing infants and children to neuro-atypical individuals and the aging population [9]. However, there are two main challenges that arise here: first, assessing if sung speech can give measures of cognition comparable to speech, and second, understanding the impact of melody on such cognition. Indeed, this is a unique scenario involving strong lyrics and melodies integration, and its perception is likely to rely on a complex interplay of speech and music processing and neural mechanisms specific to their integration.

Despite their differences, there is evidence that speech and music auditory perception likely rely on similar processes [10]. Previous work largely focused on identifying shared and distinct cortical areas responsible for the processing of speech and music with and without lyrics with technologies with high spatial resolution, such as fMRI [11, 12]. However, spatial overlap or segregation does not inform us whether speech, music, and sung speech processing employ a similar brain mechanism, *i.e.*, activation in different brain areas may rely on the same type of hierarchical statistical learning [13, 14] or sensory-motor principle [15] and therefore might give comparable measures of cognition. Previous studies investigating either speech or music encoding using non-invasive neurophysiology provided evidence in this sense. Cortical signals were shown to track progressively more abstract properties at different time scales, from

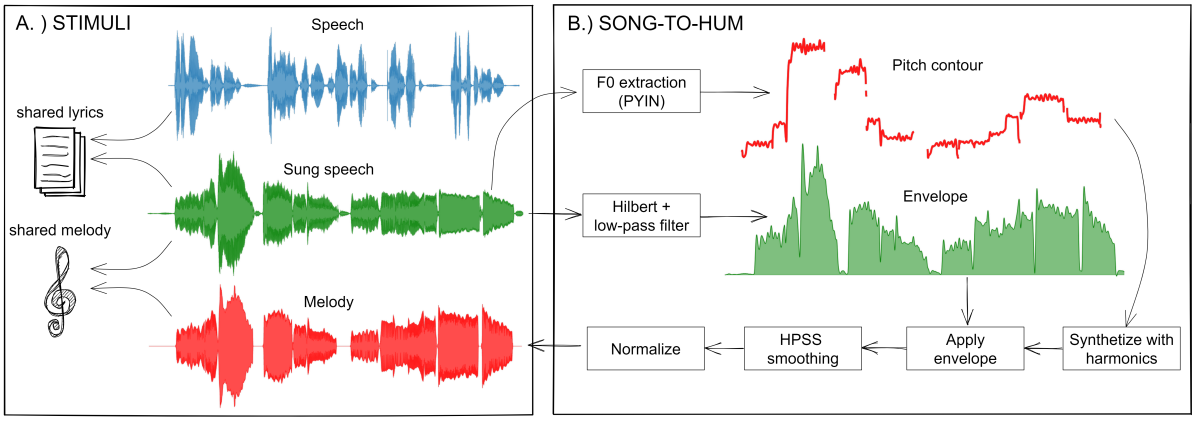

Figure 2: *(A) The set of stimuli was composed of speech, monophonic music, and sung speech. The spoken and sung versions share linguistic content. The sung and music-only versions share the melody. (B) Eighteen lyrics from the NHSS Speech and Singing Parallel Database [7] recorded by the same singer were used. Melodies were synthesized based on an ad-hoc song-to-hum algorithm.*

sound acoustics and phonological units to semantics for speech, and build predictions of upcoming sensory events, making perception an active process [16, 14]. In the case of music, this progression can be recast as fine-grained acoustic features, musical units (notes), and features at longer time scales, such as the melodic structure [17, 13]. However, it remains unclear how speech and music processing compare because of the intrinsic differences between the two hierarchies.

On the decoding side, while stimulus reconstruction is also successful for music [18, 19, 20], the reconstruction quality tends to be lower if compared to speech, even for simple acoustic features such as energy fluctuations, but direct comparisons are confounded by differences in the stimuli spectra [21, 22]. There is a consensus among researchers that the brain tracks spectro-temporal energy fluctuations for both speech and music, a phenomenon that reflects acoustic processing and encoding of higher-order features and cognitive states [17, 21]. However, it remains unclear whether the differences in the processing of speech and music are due to different linguistic and non-linguistic demands or distinct acoustic features [21]. For example, the broadband envelope of polyphonic music with instruments playing in different registers and with percussive components (*e.g.*, rock music) cannot be as much informative as the speech envelope because it will be almost flat, while polyphonic music benefits from a finer audio representation that highlights modulations in different frequency bands [18]. Therefore, previous comparative studies have yet to answer these questions because this confound cannot be eliminated without an ad-hoc experimental protocol for the comparison. Here, we propose a novel paradigm to systematically evaluate the relationship of cortical tracking of sung speech with that of speech and music alone, offering a benchmark for using sung speech in auditory research with ecologically-valid tasks and assessing spoken language and music processing abilities in a single experiment.

2. Methods

Here we propose a novel protocol for comparing music and speech cortical tracking in ecologically-valid scenarios using neural responses collected from the same set of subjects and stimuli designed ad-hoc for comparison. Specifically, we propose to use sung speech, a unique case study that implies a simultaneous encoding of speech and music, and compare its encoding to the one of music and speech only. For this purpose, we collected EEG responses to parallel monophonic music (melodies), speech, and sung speech. The set of stimuli was explicitly built in such a way as to have the most direct possible comparison among the three conditions and reduce possible confounding factors. Therefore, the linguistic content is shared across spoken and sung speech but is not rhythmic or pitched in the spoken speech stimuli, as can be seen in panel A of Figure 2. Conversely, music stimuli share the same melodic content as sung speech stimuli, but the lyrics are not pronounced.

2.1. Stimuli creation

We used the *NHSS Speech and Singing Parallel Database* [7], which collects a set of lyrics available in the sung and spoken version recorded by the same singer at 44.1 kHz along with the phoneme-level alignment, which can be used to study the encoding of higher-level speech features like phoneme-level encoding and potentially semantics. To ensure a fair comparison, pairs of conditions should include the same amount of information content (either melodic or linguistic). Therefore, we selected 18 lyrics with a minimum of 20 minutes of spoken content, corresponding to roughly 43 minutes of sung speech and the same amount for melodies. The corresponding melody was extracted from the sung speech waveform using `pYIN` [23], a modification of the `YIN` algorithm for fundamental frequency (F0) estimation [24]. In the first step of `pYIN`, F0 candidates and their probabilities are computed using the `YIN` algorithm. In the second step, Viterbi decoding is used to estimate the most likely F0 sequence and voicing flags. We limited the possible F0 candidates in the range C2($\sim$65 Hz)-C7 ($\sim$2093 Hz), and set the frame and hop size for the STFT used for the pitch extraction to respectively 2048 and 128 samples at 44.1 kHz. The aim was to create a "hummed" version of the melody without any specific instrument timbre, especially those with sharp attacks like the piano. Previous research has shown that the brain is highly sensitive to percussive note onsets (characterized by an energy burst in the spectrogram), and we wanted to avoid this possible cofound. To this end, we synthesized it using the extracted F0 and $n = 2$ harmonics, with a fade in/out and frequency interpolation transition length of 10 ms using `MeloSynth` [25]. Finally, to better mimic the human voice fluctuations, we applied

the energy envelope of the original signal and smoothed sharp attacks using a median-filtering on the spectrogram [26, 27].

2.2. Experimental procedure

Each participant listened to the same set of stimuli but in a different randomized order. To prevent participants from listening to different versions (melody, speech, and sung speech) of the same song sequentially, we divided the trials into three main experimental blocks, where each subject listened to at least one version of each song. We included two main pauses between experimental blocks, during which the subjects could take a break for 10-15 minutes. However, they were allowed to take short breaks between trials if necessary. To assess engagement with the stimuli, we included the following behavioral task: at the end of 20% of the trials, participants were asked a question about the stimulus they had just heard. For speech trials, the question was about the content, while for music trials, it was whether a little melody excerpt had been extracted from the stimulus they had just heard. For sung speech, the question could be one of two types, with a 50% chance of each.

2.3. Subjects and data acquisition

The study was conducted in accordance with the Declaration of Helsinki and was approved by the CERES committee of Paris Descartes University. Sixteen healthy individuals participated in the study (7 females, aged 23 to 51, mean age 28, with 3 left-handed individuals) who had no history of hearing impairment or neurological disorders. All except one were native English speakers. Written informed consent was obtained from all participants, who were compensated for their participation. Each participant was tested in a single session and completed a general demographic test and the Goldsmiths Musical Sophistication test (Gold-MSI) to assess individual differences in musical listening abilities [28]. Participants listened to audio samples while sitting in a sound-proof, electrically shielded booth in dim light conditions. Audio stimuli were presented monophonically at a sampling rate of $44.1\,\mathrm{kHz}$ using Sennheiser HD650 headphones and `Psychopy` [29] customized Python code. Subjects were instructed to maintain visual fixation on a crosshair centered on a screen and minimize motor activities. 64-channel EEG data and two extra electrodes on the mastoid bones were recorded and digitized at $2048\,\mathrm{Hz}$ using a BioSemi Active Two system. A customized analog system ensured optimal synchronization between the stimulus and the EEG responses.

2.4. EEG preprocessing

EEG data were analyzed offline using `MNE Python` [30, 31] and Matlab. The preprocessing pipeline follows guidelines provided for the linear modeling of neurophysiological data to auditory continuous stimuli [2]. First, the EEG was segmented into trials at the original sampling rate to avoid synchronization issues. Secondly, each trial was filtered between 1 and $30\,\mathrm{Hz}$ using low- and high-pass Butterworth zero-phase filters (order 3 with a forward and backward pass) and, finally, downsampled to $64\,\mathrm{Hz}$. Channels with a variance exceeding three times that of the surrounding ones were replaced by an estimate calculated using spherical spline interpolation. All channels were re-referenced to the average of the two mastoid channels to maximize auditory responses [2]. The first $500\,\mathrm{ms}$ at the start of each trial were removed to avoid responses elicited by the start of the stimulus. Finally, EEG responses of each subject were standardized together to preserve the relative power across channels.

2.5. Encoding model

The broadband amplitude envelope was extracted from the acoustic waveforms using the Hilbert transform and downsampled to $64\,\mathrm{Hz}$. Then, the channel-specific mapping between the amplitude envelope and the neural data, namely the TRF, was estimated by solving a regularized linear regression problem [6]. How well the model predicts unseen data is quantified by Pearson's correlation (r) between the predictions and the real neural recording for each channel using a leave-one-out cross-validation procedure. The model is learned considering multiple stimulus-response time lags τ. We first fit the model using an exploratory time-lag window ranging from -100 to $600\,\mathrm{ms}$, which was later restricted to the significant peaks between 0 and $300\,\mathrm{ms}$ for the final analysis. Note that when training, the considered time window is always $[\tau_{min} - \delta, \tau_{max} + \delta]$, where δ =$50\,\mathrm{ms}$ to avoid border artifacts. While forward models offer a view of the spatiotemporal dynamics of the cortical responses to a stimulus, backward models combine multivariate and noisy EEG data, thus reducing redundancy and autocorrelation, to reconstruct the univariate and clean sound envelope, leading to a larger and more reliable correlation score [6]. Here, we fit a decoding model using τ_{min} =0 and τ_{max} =$300\,\mathrm{ms}$.

3. Results

Our goal is to quantify how well neural responses track sound energy fluctuations in the three different conditions, and investigate if there is an effect of melody on the tracking of sung speech with respect to speech-only. We used forward and backward models to describe how the acoustic envelope is transformed into EEG signals and vice versa in the three conditions.

First, forward encoding models were fitted for each condition, subject, and EEG channel. The resulting TRFs were then averaged within each condition to explore the system's temporal dynamics in response to music-only, speech-only, and sung speech. Figure 3A displays the TRFs of all channels for the three conditions to highlight dipoles occurring at approximately 60, 125 , and $200\,\mathrm{ms}$ corresponding to a negative deflection N1, a positive peak P1 of the centro-parietal area, and a second, but weaker, negative deflection N2, particularly for the speech condition (as confirmed by the global field power, *i.e.*, the standard deviation of the TRF weights across channels). Below, the TRF weights for each condition and peak of interest are shown on topographical maps. We found significant differences in the amplitude of N1 for the sung speech/speech pair and in the degree of polarization for N1 in the sung speech/melody pair ($p < 0.05$, t-test, FDR correction), as can be seen on the topographies in Figure 3B. The weight differences can be interpreted as the contribution of lyrics and melody in sung speech when looking respectively at the sung speech/melody and sung speech/speech differences. Interestingly, at lower frequencies (1-8 Hz), the significant difference between melody and sung speech disappears. In contrast, the one in N1 between sung speech and speech becomes clearer, with significantly higher weights associated with the frontal electrodes ($p < 0.05$, t-test, FDR correction), indicating an effect of melody processing when listening to sung speech. To further investigate this effect in low frequencies, we analyzed the TRFs of channels of interest. N1, P1, and N2 are significantly larger than zero for Fz and Cz, while for Pz only P1 was significant ($p < 0.05$, t-test). The effect size of the condition amplitude difference was calculated at each time lag, showing a medium effect of speech encoding corresponding to the N1 peak for centro-frontal electrodes (Co-

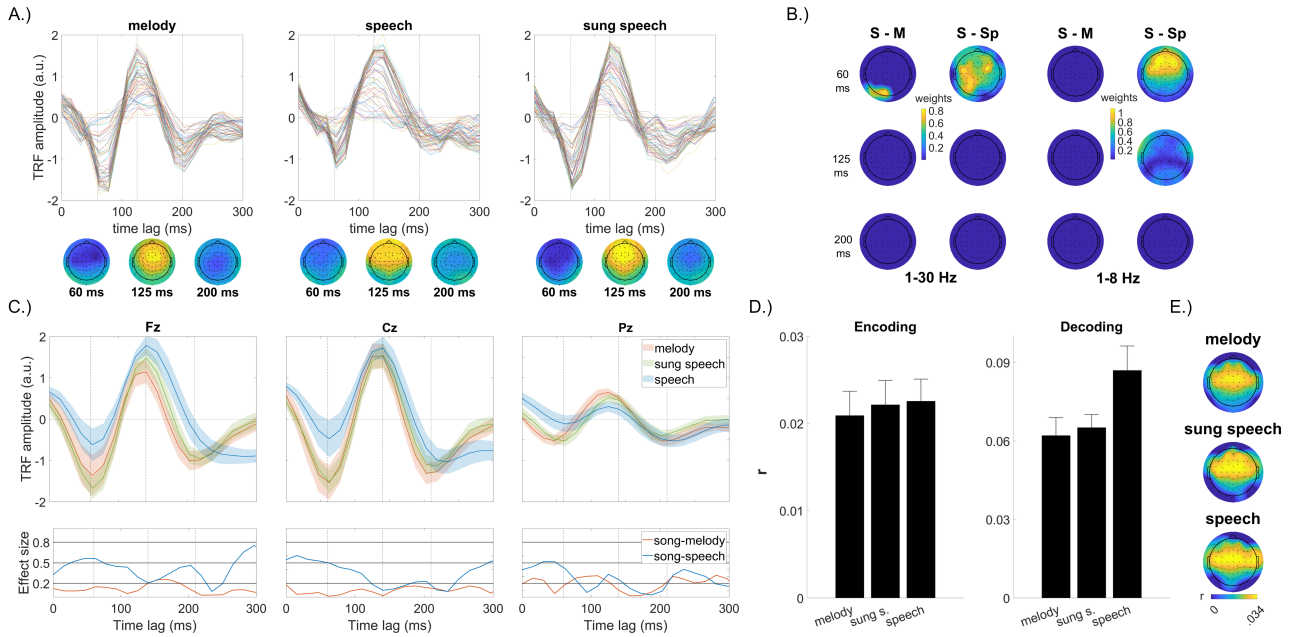

Figure 3: *(A) TRFs of all channels and conditions to highlight dipoles together with TRF weights for each peak of interest on topographical maps. (B) Difference between TRF weights (absolute value) in pairs of conditions (sung speech/melody, sung speech/speech) at each peak of interest and different filtering. Only significant channels are displayed ($p < 0.05$, t-test, FDR correction). (C) TRFs at Fz, Cz, and Pz (mean±standard error across subjects) considering a* 1-8 Hz *filtering together with the effect size computed as Cohen's d between pairs of conditions. (D) Encoding (average over all channels) and decoding correlations (mean±standard error across subjects and trials). Both analyses show significant correlations for all conditions ($p < 0.001$, t-test against the null distribution) but no effect of the condition ($p > 0.05$, one-way ANOVA, Bonferroni correction). (E) EEG prediction correlations displayed on the scalp topography. Only significant channels ($p < 0.05$, t-test, Bonferroni correction) are displayed.*

hen's d > 0.5, Figure 3C). Next, the EEG data were predicted using the TRF models. The distribution of correlation scores was significantly better than the null distribution obtained with shuffled features for all the conditions ($p < 0.001$, *t-test*), with no significant difference across conditions ($p > 0.05$, t-test, Bonferroni correction, Figure 3D). The best-predicted electrodes were located in a broad centro-parietal area of the scalp for all conditions ($p < 0.05$, t-test, Bonferroni correction, Figure 3E), but no significant difference was found across conditions. Finally, to support our findings, we performed a decoding analysis. Once again, the correlation scores were significantly better than the null distribution, with no significant difference across conditions ($p > 0.05$, t-test, Bonferroni correction).

4. Conclusions

Envelope tracking has been shown to relate to various cognitive functions and can provide objective metrics of typical auditory and language processing, as well as cognitive and sensory deficits [8]. While previous research focused on speech, recent applied work highlighted the need for more engaging stimuli, such as sung speech, when considering particular cohorts, from typically developing infants to elderly and neuro-atypical individuals [9]. However, no study assesses whether sung speech can give measures of cognition comparable to speech and analyzes the impact that melody might have. We propose a novel paradigm that involves sung speech, hummed melodies, and spoken lyrics, specifically designed and built for the study and comparison of spoken language and music encoding in a single experiment while minimizing possible confounds due to different envelope spectra [21]. Previous work has compared envelope tracking of speech and polyphonic music, which, however, is significantly more complex in terms of time-frequency modulation [18], making the envelope a poor descriptor for music and the comparison unfair. Indeed the envelope is not sufficiently informative for polyphonic music, but it is a more sensible descriptor when considering simple monophonic melodies.

Here, we provided the first direct comparison of envelope tracking of sung speech, speech, and melody, offering a benchmark for using sung speech in auditory research with ecologically-valid tasks and encouraging the use of sung speech in applied research. Indeed, our analyses using encoding and decoding modeling showed robust and consistent envelope tracking across spoken and sung speech, even when considering lower frequencies in the EEG. The only significant differences were exclusively due to melody processing, indicating that sung speech can give measures of cognition comparable to those obtained with speech. We also assessed the impact of melody on the linguistic processing of sung speech, finding enhanced activity corresponding to N1 in the centro-frontal area.

While in this work we only considered envelope tracking, this framework will ultimately allow studying higher-order speech and music features (*e.g.* semantics, melodic structure), potentially leading to a variety of neural indices of speech and music processing at various levels of abstraction.

5. Acknowledgements

This paper used the NHSS Database made available by the HLT lab, National University of Singapore.

6. References

[1] L. S. Hamilton and A. G. Huth, "The revolution will not be controlled: natural stimuli in speech neuroscience," *Language, cognition and neuroscience*, vol. 35, no. 5, pp. 573–582, 2020.

[2] M. J. Crosse, N. J. Zuk, G. M. Di Liberto, A. R. Nidiffer, S. Molholm, and E. C. Lalor, "Linear modeling of neurophysiological responses to speech and other continuous stimuli: methodological considerations for applied research," *Frontiers in Neuroscience*, p. 1350, 2021.

[3] J. Obleser and C. Kayser, "Neural entrainment and attentional selection in the listening brain," *Trends in cognitive sciences*, vol. 23, no. 11, pp. 913–926, 2019.

[4] N. Ding and J. Z. Simon, "Cortical entrainment to continuous speech: functional roles and interpretations," *Frontiers in human neuroscience*, vol. 8, p. 311, 2014.

[5] J. A. O'sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial eeg," *Cerebral cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[6] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli," *Frontiers in human neuroscience*, vol. 10, p. 604, 2016.

[7] B. Sharma, X. Gao, K. Vijayan, X. Tian, and H. Li, "Nhss: A speech and singing parallel database," *Speech Communication*, vol. 133, pp. 9–22, 2021.

[8] J. Palana, S. Schwartz, and H. Tager-Flusberg, "Evaluating the use of cortical entrainment to measure atypical speech processing: a systematic review," *Neuroscience & Biobehavioral Reviews*, vol. 133, p. 104506, 2022.

[9] A. Attaheri, D. Panayiotou, A. Phillips, Á. N. Choisdealbha, G. M. Di Liberto, S. Rocha, P. Brusini, N. Mead, S. Flanagan, H. Olawole-Scott *et al.*, "Cortical tracking of sung speech in adults vs infants: A developmental analysis," *Frontiers in Neuroscience*, vol. 16, 2022.

[10] L. S. Hamilton, "Human song: Separate neural pathways for melody and speech," *Current Biology*, vol. 32, no. 7, pp. R311–R313, 2022.

[11] P. Albouy, L. Benjamin, B. Morillon, and R. J. Zatorre, "Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody," *Science*, vol. 367, no. 6481, pp. 1043–1047, 2020.

[12] S. V. Norman-Haignere, J. Feather, D. Boebinger, P. Brunner, A. Ritaccio, J. H. McDermott, G. Schalk, and N. Kanwisher, "A neural population selective for song in human auditory cortex," *Current Biology*, vol. 32, no. 7, pp. 1470–1484, 2022.

[13] G. M. Di Liberto, C. Pelofi, R. Bianco, P. Patel, A. D. Mehta, J. L. Herrero, A. de Cheveigné, S. Shamma, and N. Mesgarani, "Cortical encoding of melodic expectations in human temporal cortex," *Elife*, vol. 9, p. e51784, 2020.

[14] C. Caucheteux, A. Gramfort, and J.-R. King, "Evidence of a predictive coding hierarchy in the human brain listening to speech," *Nature Human Behaviour*, pp. 1–12, 2023.

[15] S. Shamma, P. Patel, S. Mukherjee, G. Marion, B. Khalighinejad, C. Han, J. Herrero, S. Bickel, A. Mehta, and N. Mesgarani, "Learning speech production and perception through sensorimotor interactions," *Cerebral cortex communications*, vol. 2, no. 1, p. tgaa091, 2021.

[16] C. Brodbeck, L. E. Hong, and J. Z. Simon, "Rapid transformation from auditory to linguistic representations of continuous speech," *Current Biology*, vol. 28, no. 24, pp. 3976–3983, 2018.

[17] G. M. Di Liberto, C. Pelofi, S. Shamma, and A. de Cheveigné, "Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 361–364, 2020.

[18] G. Cantisani, S. Essid, and G. Richard, "Eeg-based decoding of auditory attention to a target instrument in polyphonic music," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 80–84.

[19] G. M. Di Liberto, G. Marion, and S. A. Shamma, "Accurate decoding of imagined and heard melodies," *Frontiers in Neuroscience*, vol. 15, p. 673401, 2021.

[20] L. Hausfeld, N. R. Disbergen, G. Valente, R. J. Zatorre, and E. Formisano, "Modulating cortical instrument representations during auditory stream segregation and integration with polyphonic music," *Frontiers in neuroscience*, vol. 15, p. 635937, 2021.

[21] N. J. Zuk, J. W. Murphy, R. B. Reilly, and E. C. Lalor, "Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies," *PLoS computational biology*, vol. 17, no. 9, p. e1009358, 2021.

[22] A. M. D. Simon, J. Østergaard, S. Bech, and G. S. J. M. Loquet, "Optimal time lags for linear cortical auditory attention detection: differences between speech and music listening," in *Proc. Int. Symposium on Hearing*, 2022.

[23] M. Mauch and S. Dixon, "pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 IEEE Int. Conf. on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 659–663.

[24] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[25] J. Salomon. (2018) Melosynth. [Online]. Available: https://github.com/justinsalamon/melosynth

[26] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, vol. 13, 2010, pp. 1–4.

[27] J. Driedger, M. Müller, and S. Disch, "Extending harmonic-percussive separation of audio signals." in *International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 611–616.

[28] D. Müllensiefen, B. Gingras, J. Musil, and L. Stewart, "Measuring the facets of musicality: the goldsmiths musical sophistication index (gold-msi)," *Personality and Individual Differences*, vol. 60, p. S35, 2014.

[29] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, "Psychopy2: Experiments in behavior made easy," *Behavior research methods*, vol. 51, pp. 195–203, 2019.

[30] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen *et al.*, "Meg and eeg data analysis with mne-python," *Frontiers in neuroscience*, p. 267, 2013.

[31] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and M. S. Hämäläinen, "Mne software for processing meg and eeg data," *Neuroimage*, vol. 86, pp. 446–460, 2014.