Auditory Cortex Tracks Masked Acoustic Onsets in Background Speech: A Potential Stream Segregation Mechanism

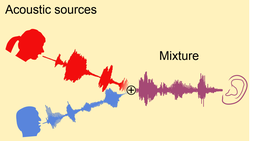
Computational Sensorimotor Systems Lab

Christian Brodbeck[1]*, Alex Jiao[1], L. Elliot Hong[2] & Jonathan Z. Simon[1]

[1]University of Maryland, College Park; [2]University of Maryland School of Medicine; *christianbrodbeck@me.edu

Motivation

- Listening to speech in the presence of multiple talkers
- The acoustic stimulus is an additive mixture of multiple speech waveforms (here: monophonic presentation)



- Listeners need to segregate features of the attended speaker
- Previous work shows early neural representation of the acoustic mixture (~50 ms latency) and later representation of attended speaker (~100 ms) (Puvvada and Simon 2017; O'Sullivan et al. 2019)
- **Are early representations restricted to passive spectro-temporal filtering of the mixture, or do they also involve active extraction of acoustic features?** To what degree are such features actively segregated and represented as auditory objects?
- **Acoustic onsets**:
  - Important for auditory object formation and, consequently, stream segregation
  - Simultaneous onsets in multiple frequency bands indicate that the different spectro-temporal elements have a common physical source

For details see preprint: https://doi.org/10.1101/866749

Methods

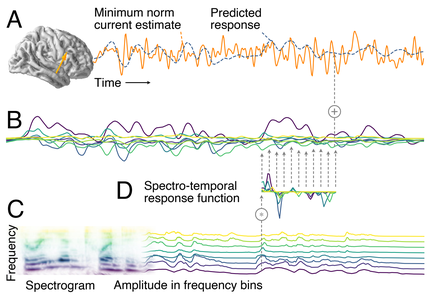
Participants listened to 1 minute long **audiobook segments** in two conditions:

- Single talker
- Two talkers: one male, one female;
  - Task: attend to one speaker, ignore the other
  - Attention counterbalanced across trials and participants

Whole head magnetoencephalography (**MEG**)

- Localized to cortical surface (minimum norm estimates)

Brain responses were modeled as linear convolution of predictor variables representing the stimuli with to-be-estimated **temporal response functions (TRFs)**.



A) Sample response in one current dipole. Model fit evaluated as the Pearson correlation between measured and predicted responses. **B)** The predicted response was the sum of the responses to different predictor time series, modeling concurrent brain responses to different features of the input. **C)** For model estimation, spectrograms were decomposed into 8 frequency bins. **D)** Multi-dimensional kernels, estimated with a coordinate descent algorithm. Quantify response to different stimulus frequencies: spectro-temporal response functions (STRFs).

Model comparisons to evaluate the contribution of each predictor to the model fit. For each predictor, the model fit of full model was compared to a model in which this predictor was temporally misaligned with the MEG responses.

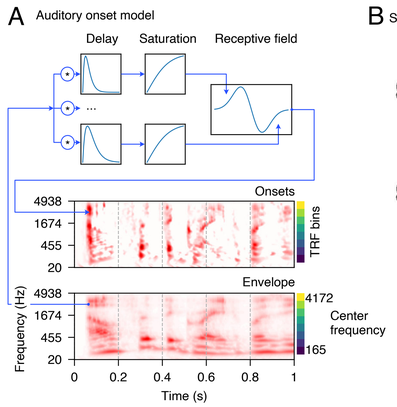
Results

1. Single talker

Single speaker reading audiobook excerpts

A) Predictor variables

- Onsets: acoustic onsets, extracted from the gammatone spectrogram, using a neurally inspired edge detector (Fishbach, Nelken, and Yeshurun 2001)
- Envelope: sustained acoustic signal from the gammatone spectrogram



B) Model fit ($p \leq .05$, corrected)

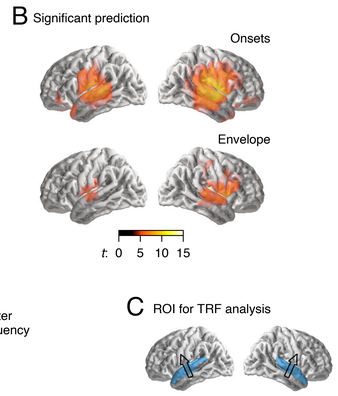
- Both predictors contributed to brain responses, localization consistent with sources in superior temporal gyrus (STG)

C) Region of interest (ROI)

- Temporal response functions were analyzed in STG, positive values for upward current

D) Temporal response functions (TRFs)

- Onsets: strong upward peak (~70 ms latency) followed by downward peak (~130 ms)
- Envelopes: TRFs are diminished compared to acoustic onsets



2. Two talkers

Two talkers, male/female, equal loudness

A) Predictors

- Acoustic onsets and envelopes each for:
  - The acoustic mixture (heard by participants)
  - The unmixed to-be attended speaker
  - The unmixed to-be ignored speaker

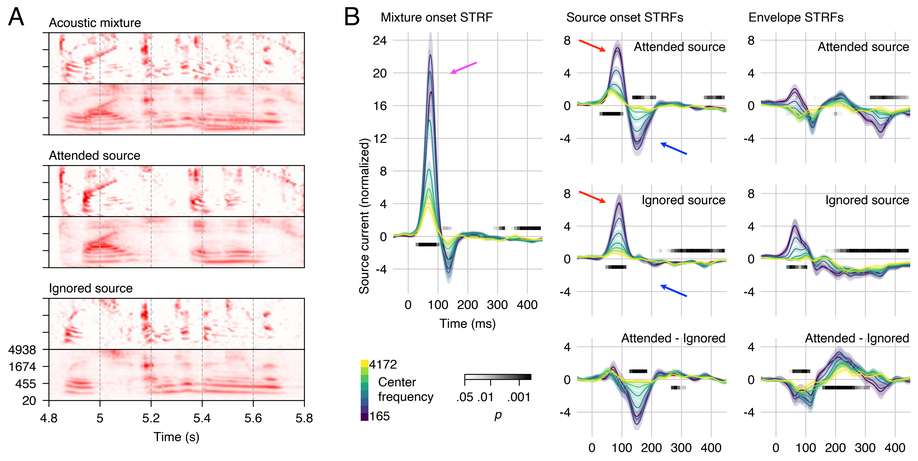
Model fits

- Onsets: significant representation of onsets in the ignored speaker even after controlling for the acoustic mixture and the attended speaker
- Envelope: The ignored source could not be statistically distinguished from a linear combination of the mixture and the attended source

B) Temporal response functions

Onsets:

- →Early response to onsets in the acoustic mixture
- →Additional, early response to onsets in either of the sources; **suggests that onsets in both speakers are initially recovered, even if they are not overtly present in the mixture**
- →Later, negative response to onsets only in the attended source



3. Masked onsets

A) The two-talker results (2) suggest that onsets in either speaker are represented, even if they are masked in the mixture by the other source. To further explore this, overt and masked onsets were modeled separately for each of the speakers.

- →Overt onsets: occur in one of the speakers, and are also apparent as onsets in the acoustic mixture
- →Masked onsets: occur in one of the speakers at times where there is no corresponding onset apparent in the mixture

Model fits: even masked onsets in the ignored speaker significantly improve model fit

B) Temporal response functions (TRFs)

Early, upward peak: mainly bottom-up response

- →Larger response to overt compared to masked onsets
- →Small effect of attention on masked onsets only

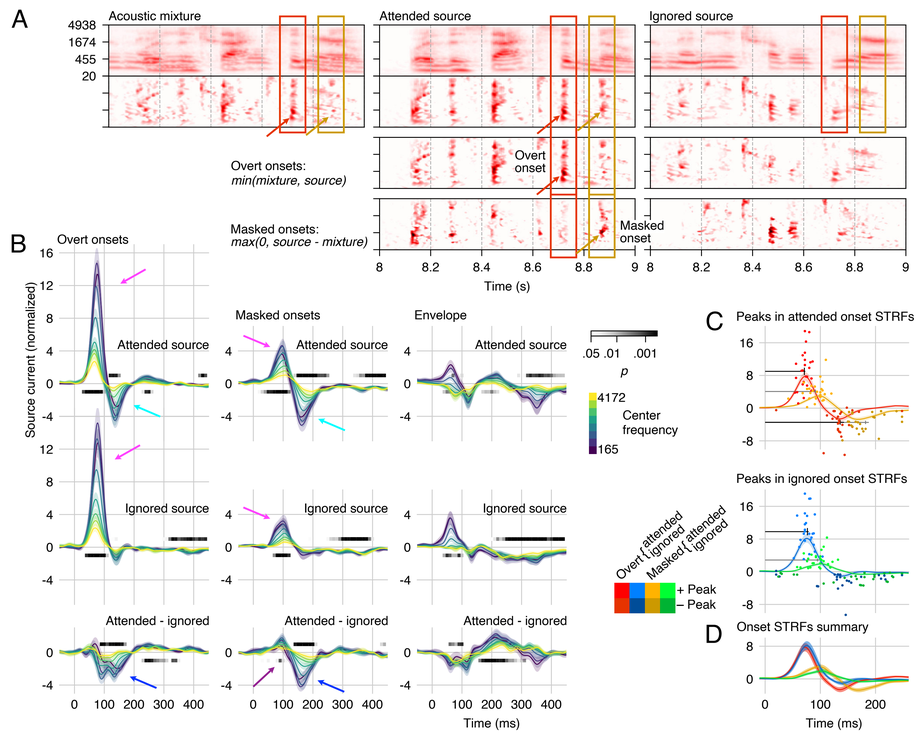
Later, downward peak: attentional processing

- →Amplitude similar for overt and masked onsets
- →Strong effect of attention

C) Peak latency analysis

- Both early and later peaks occur significantly later for masked onsets compared to overt onsets

Onsets in ignored speech are not just passively perceived when they are overt in the acoustic signal, but are represented even when they are masked, with a temporal processing cost



Discussion

Main result: Acoustic features (onsets) in the ignored speaker are represented in auditory cortex even if they are not apparent in the acoustic mixture

- Suggests reconstruction of features that are masked in the input, neural "filling in"
- Suggests auditory object representations, including (small) attentional influence, even in early responses

Active segregation of features of the ignored speech could explain behavioral results:

- Speech comprehension in the presence of another talker is harder than in the presence of spectrally matched noise
- In multi-speaker environment, unintentional switching to unattended speaker is more likely than simple inability to understand attended speaker
- Auditory (proto-) objects of the ignored speaker could explain attentional capture and bottom-up switching to ignored speaker

References

Fishbach, Alon, Israel Nelken, and Yehezkel Yeshurun. 2001. "Auditory Edge Detection: A Neural Model for Physiological and Psychoacoustical Responses to Amplitude Transients." Journal of Neurophysiology 85(6):2303–23.

O'Sullivan, James, Jose Herrero, Elliot Smith, Catherine Schevon, Guy M. McKhann, Sameer A. Sheth, Ashesh D. Mehta, and Nima Mesgarani. 2019. "Hierarchical Encoding of Attended Auditory Objects in Multi-Talker Speech Perception." Neuron S0896627319307809.

Puvvada, Krishna C., and Jonathan Z. Simon. 2017. "Cortical Representations of Speech in a Multitalker Auditory Scene." Journal of Neuroscience 37(38):9189–96.