Automated Word Timing from Speech Audio for Brain Signal Analysis

Stella Alumonah, Vrishab Commuri, Charlie Fisher, Jonathan Z. Simon
Department of Electrical and Computer Engineering

Computational Auditory Neural Systems Laboratory
UMD-REACH
Research, Equity, and Access in Communication and Hearing

Introduction

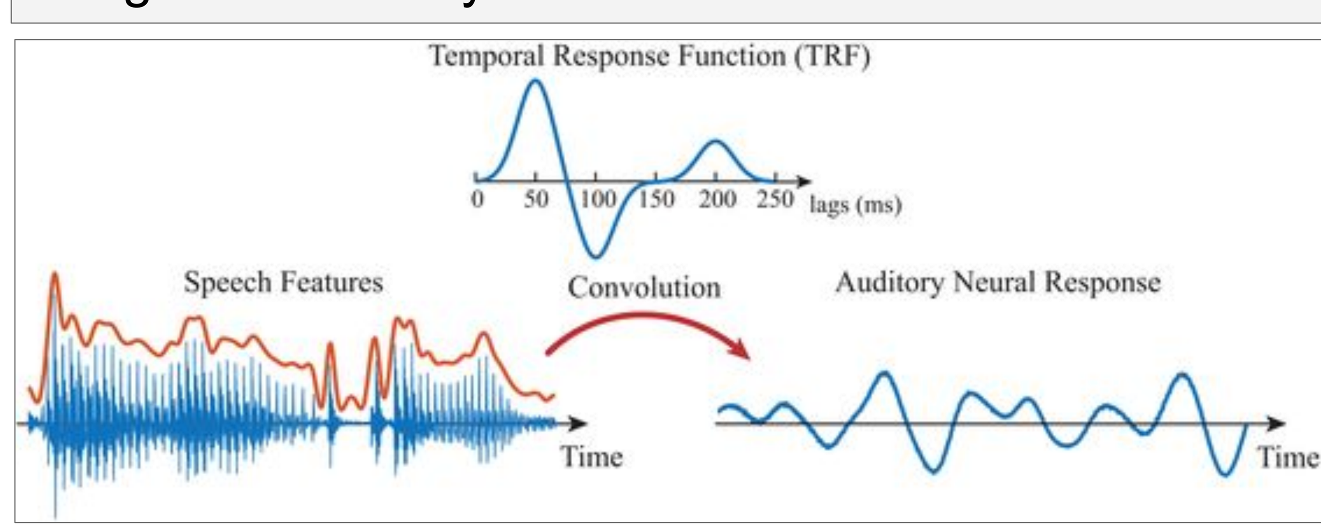
- When the brain processes speech, it actively identifies and interprets words.
- During speech listening, the brain imaging technique magnetoencephalography (MEG) measures magnetic signals generated by brain electrical activity
  - Including signals identifying and interpreting words.
- In a typical experiment, subjects listen to speech recordings, and MEG shows neural responses elicited by the start of each word.
  - To see this, the times at which each word is spoken must be known.

Gap: Several speech-to-text alignment tools have been designed to generate word-level timestamps from audio files, however researchers often have to manually adjust the results. With large amounts of data, this essential task becomes especially time-consuming.

Question: How can a variety of speech alignment models collaborate to automate the process of extracting word timings and improve timing accuracy?

Methods

- Extracted word-level timestamps from speech processing systems
- Twelve ~1-minute-long speech recordings from single-speaker passages of an audiobook
- Recorded MEG data from five participants
- Implemented Python algorithm to standardize the output from all models with the original transcript as reference
- Analyzed how closely the combination of four systems would estimate the human annotated timings using the scikit-learn linear regression library



Impulse response framework to predict auditory neural response

- To visualize the quality of the automated timings, Temporal Response Functions (TRFs) were constructed to map word onsets to neural responses

Evaluation

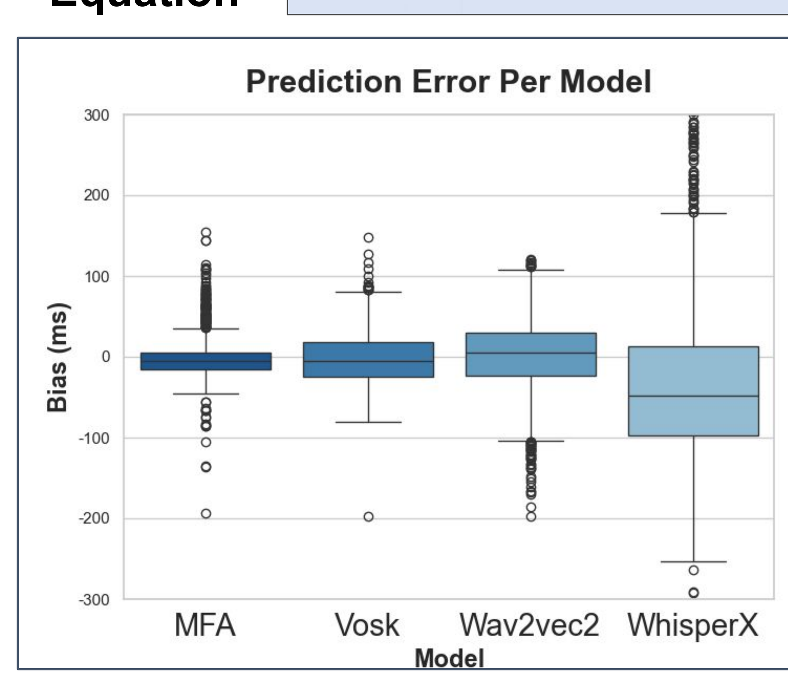
Four approaches were utilized that combine a list of classical and modern speech alignment systems.

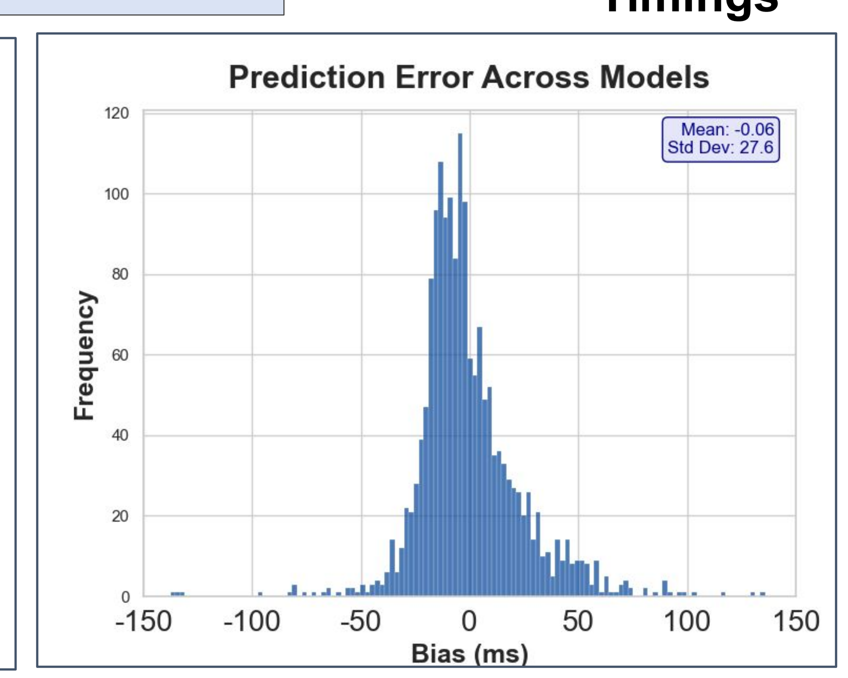
Model Type	System	Functionality
Montreal Forced Aligner (3.3.0)	Kaldi Gaussian Mixture Model – Classical Machine Learning	Time-aligns a transcript with audio using a pronunciation dictionary and acoustic model.
WhisperX (Large –v2)	Open AI Whisper with Wav2vec2 – Deep Neural Network	Transcribes 30s audio chunks; estimate word timings via phoneme recognition
Wav2vec2 (Large-960h -lv60-self)	Transformer model – Deep Neural Network	Converts audio into feature vectors; aligns transcribed text to signal
Vosk (en-us-0.22)	Kaldi Model – Deep Neural Network	Processes audio in small frames with real time text and timing prediction

Results

- Each model's regression was evaluated based on how closely it matched the human annotated timings
  - MFA (Least Variability □ Consistent Predictions)
  - Vosk (Median closest to zero □ More accurate)
  - Wav2Vec2(More errors with delayed predictions)
  - WhisperX (Greatest variability, significant outliers)

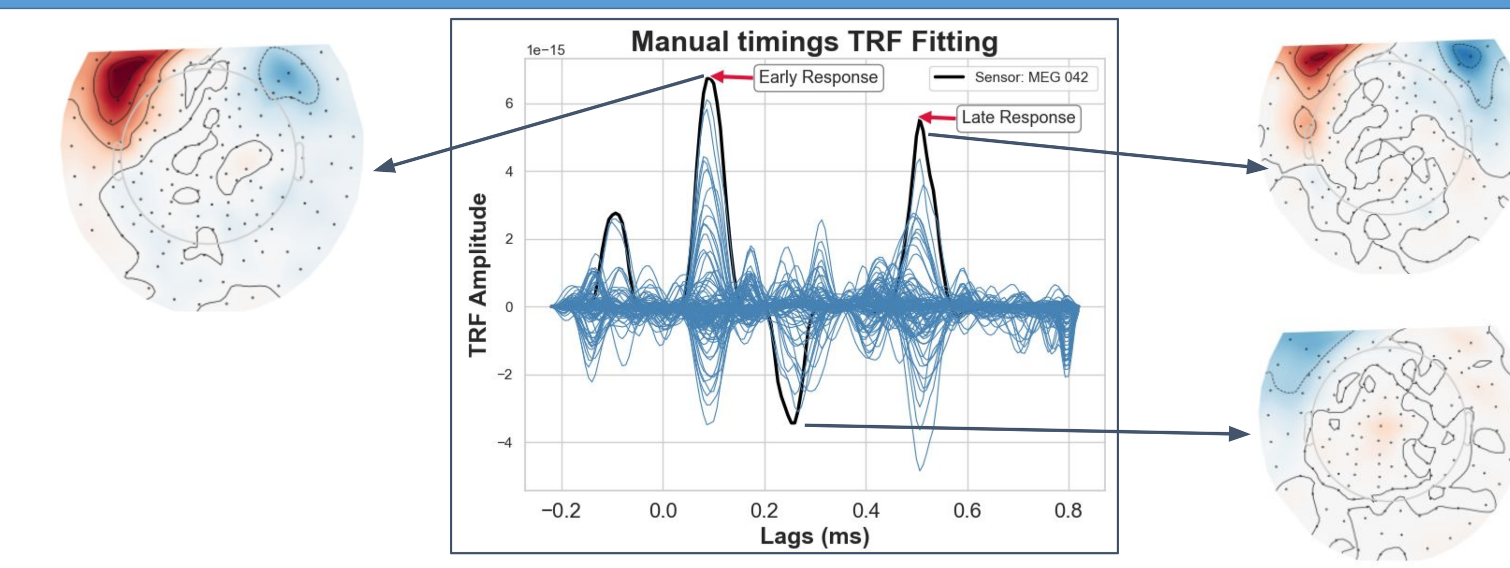
Regression Equation $\hat{y} = \beta_0 + \beta_1 X + \varepsilon$ Predicted Timings

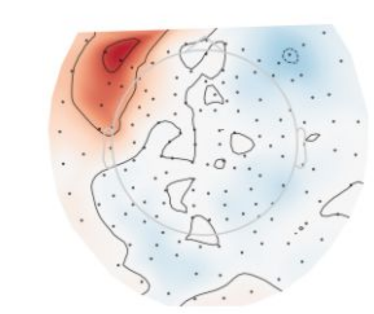


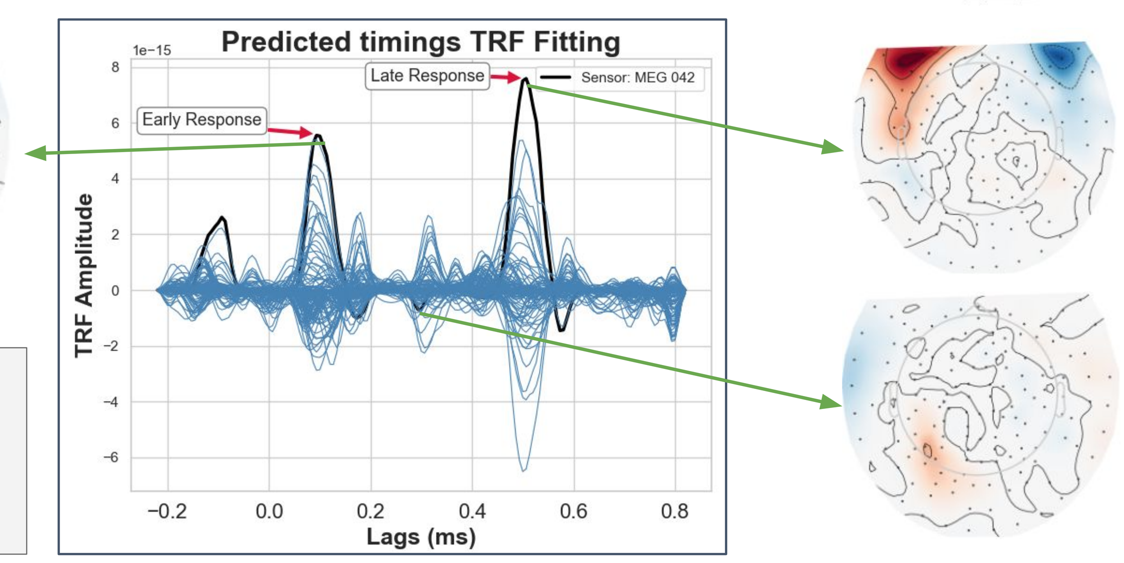


- Difference between predicted times across all models and the manually annotated timings lay between -25 to 25 ms
  - Ideally, this would be 0 confirming perfect predictions
- Right skewness indicate tendency of the models to systematically delay timings

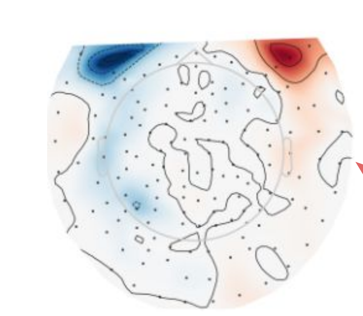
TRF Analysis

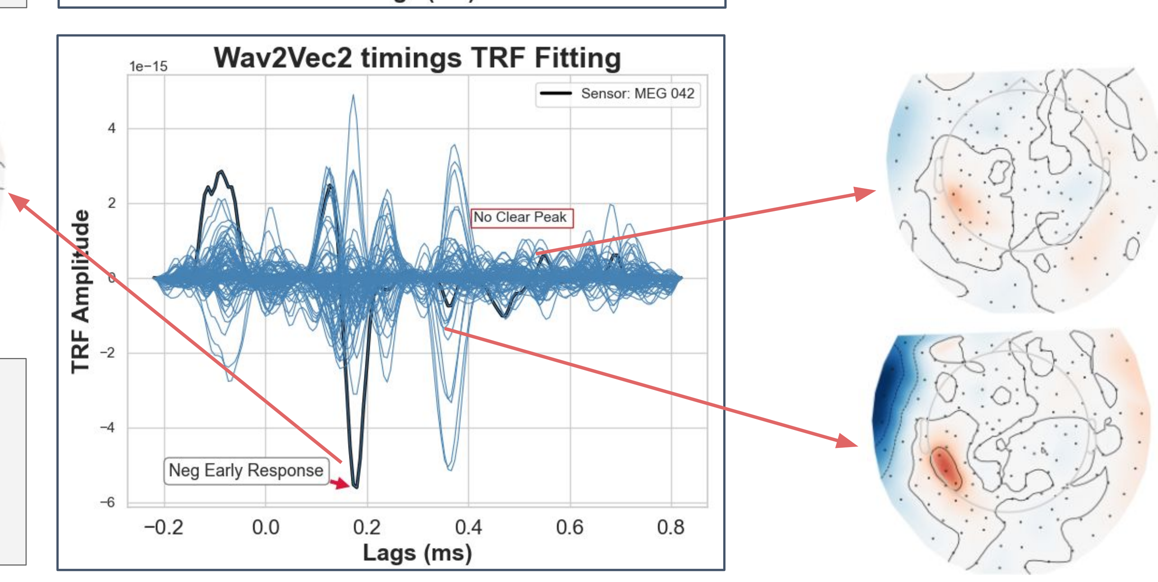






Comparable peak latencies in same direction





Opposite direction with no distinct late response

Discussion

- Predictions from the combined model aligned more closely to the manual timings compared to the state-of-the-art model, Wav2vec2
  - Topographic maps demonstrate that magnetic field amplitudes mirrors that from the manual timings
- Regression-derived timings provide a reliable level of precision suitable for many analyses
- Rapid and consistent extraction of acoustic features offers clear advantage over manual timings
  - Scalability over large datasets will save time in preprocessing
- Future work will focus on better ways of combining the models to improve accuracy

Acknowledgements & References