Decoding the Unintelligible: Neural Speech Tracking in Low Signal-to-Noise Ratios

Xiaomin He [1,2], *Vinay S Raghavan* [1,2], *Nima Mesgarani* [1,2]

[1] Department of Electrical *Engineering*, Columbia University, New York, NY, 10027, United States
[2] Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, United States

***Abstract*— Understanding speech in noisy environments is challenging for both human listeners and technology, with significant implications for hearing aid design and communication systems. Auditory attention decoding (AAD) aims to decode the attended talker from neural signals to enhance their speech and improve intelligibility. However, whether this decoding remains reliable when speech intelligibility is severely degraded in real-world listening conditions remains unclear. In this study, we investigated selective neural tracking of the attended speaker under adverse listening conditions. Using EEG recordings in a multi-talker speech perception task with varying SNR, participants' speech perception was assessed through a repeated-word detection task, while neural responses were analyzed to decode the attended talker. Despite substantial degradation in intelligibility, we found that neural tracking of attended speech persists, suggesting that the brain retains sufficient information for decoding. These findings demonstrate that even in highly challenging conditions, AAD remains feasible, offering a potential avenue for enhancing speech intelligibility in brain-informed audio technologies, such as hearing aids, that leverage AAD to improve speech perception in real-world noisy environments.**

Keywords— Speech Perception, Auditory Attention Decoding, EEG, Brain-informed Hearing Aids

I. Introduction

Humans excel at selectively attending to a single speaker in noisy environments, a skill crucial for understanding speech in the real world. This ability is supported by the brain's capacity to separate concurrent sounds and enhance the target speech [1], [2], [3]. Recent advances in neuroscience have demonstrated that neural tracking of speech features, such as the speech envelope, reflects this selective attention [4], [5], [6]. Through these neural responses, auditory attention decoding (AAD) has emerged as a promising technique, enabling the identification of the attended speaker based on brain activity in challenging multi-talker environments [7] which can be used to enhance that talker.

However, a key limitation of AAD is its potential breakdown under degraded listening conditions, particularly when the target speech is difficult to perceive. The challenge lies in determining whether neural tracking—and thus AAD—remains effective when speech intelligibility is severely compromised by noise. This issue is especially critical for real-world applications, such as hearing aids, which are often needed most in noisy environments where the target speaker is difficult to discern. Previous studies have suggested that attention and intelligibility can influence the degree of neural tracking, but the specific conditions under which perceptual breakdown renders AAD ineffective remain underexplored [8], [9], [10], [11].

In this study, we address this gap by investigating the robustness of neural tracking in situations where speech intelligibility is significantly compromised. Specifically, we designed a multi-talker speech perception experiment with varying levels of masking speaker and background noise, and measured participants' speech perception using a repeated-word detection task [12], [13]. Simultaneously, we recorded EEG data to assess neural responses to both target and non-target speech using canonical correlation analysis (CCA) [14], [15]. Our results reveal that selective neural tracking of the attended speaker persists even under highly degraded conditions, suggesting that the brain retains sufficient information to decode the attended speaker, regardless of significant intelligibility loss.

This work not only advances our understanding of the neural mechanisms underlying speech perception in noise but also provides new insights into the potential of AAD to improve listening in real-world noisy environments.

II. Methods

A. Participants

Fourteen native American English speakers (7 males; mean ± standard deviation (SD) age: 24.86 ± 4.4 years) participated in the study. All participants reported normal hearing. They were compensated with a base payment, along with a performance-based bonus determined by task performance (1-back detection accuracy).

B. Experiment

The study protocol was approved by the Institutional Review Board of Columbia University. Two loudspeakers were placed at ±45 degrees in front of the participants. Participants were

This work is funded by the National Institutes of Health (NIH-NIDCD) and a grant from Marie-Josee and Henry R. Kravis.

instructed to focus on the target speech, with its gender and direction specified on a monitor. A non-target masking speech and bi-channel naturalistic background noises were played simultaneously. All speech was in American English, synthesized by Google Text-to-Speech using four different voices (2 males and 2 females). The target speech was normalized to 65 dB, with its SNRs ranging from -12 dB to 4 dB, adjusted by modulating the loudness of the masking speech and background noise.

To assist with target speech tracking in low-SNR conditions, the masking speech and background noise faded in to the specified loudness after 3 seconds of the trial onset. These initial 3-second windows were excluded from the analysis. The experiment included 160 trials, each lasting approximately 35 seconds. During the task, participants pressed a buzzer when they detected 1-back repeated words in the target speech. Their detection accuracy determined the performance-based bonus. After every round (16 trials), participants received real-time feedback on their task performance and were asked to briefly summarize the content, ensuring engagement.

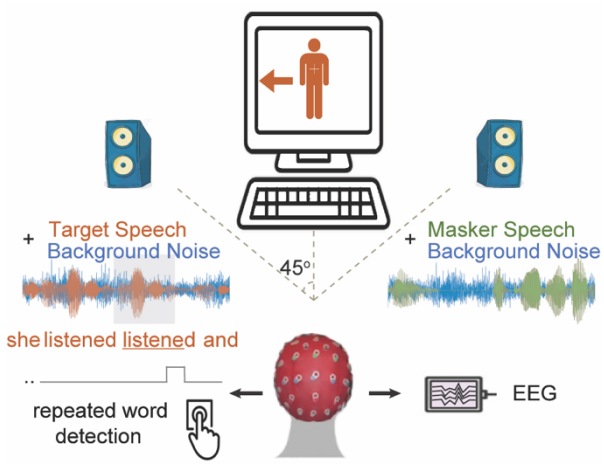

Fig. 1. **Experiment Schematic:** Participants focused on target speech amidst background speech and noises.

C. Preprocessing

For each of the 160 trials, 64-channel EEG data and buzzer responses were recorded using the g.HIAMP system (g.tec, Austria) and streamed through Simulink (MathWorks, MA, USA) at a sampling rate of 1200 Hz with a 60 Hz notch filter. EEG data were down-sampled to 100 Hz. Invalid channels with abnormal standard deviations were replaced using spherical interpolation of the remaining channels [16], [17], [18].

Speech envelopes for both target and masking speech were extracted using a nonlinear, iterative method [19] and down-sampled to 100 Hz to match the EEG recordings. The envelopes were then z-scored.

In this experiment, 1-back detection performance was used to quantify speech perception for long, continuous speech segments. The average detection accuracy for each bin of SNR, ranging from -12 dB to 4 dB (bin size, 2dB), was fitted using *psignifit* toolbox [20], [21], and a customized logistic function as in (1). The participant-specific psychometric curves, determined by the regression curve with the highest R^2 and lowest RMSE, were used to extract the average perception for each SNR bin.

$$DA = lw + \frac{lw+(up-lw)}{1+\ e^{-gr(SNR-ths)}} \tag{1}$$

where *lw* is the lower bound (approaching 0), *up* is the upper bound (approaching 1), *gr is the* growth rate, *SNR* is the signal-to-nosie ratio ranging from -12 to 4 dB, and *ths* is the threshold when DA = 50%, defined in the range of SNRs.

D. CCA Decoding Analysis

We used canonical correlation analysis (CCA) [14], [15] to evaluate the neural entrainment of speech envelope due to its simplicity and good performance in the categories of linear models [15]. The CCA model estimates the relationship between stimuli (speech envelope) and neural responses (EEG) by transforming both into a shared space that maximizes correlation. In this study, processed speech envelopes and EEG data were windowed with 200 ms and 400 ms overlapping receptive fields, respectively. The CCA model was trained using a leave-one-out cross-validation method for each participant.

For each participant and trial, the Pearson correlation between the speech envelope and EEG was computed to quantify neural speech tracking. The correlation for the target speech is denoted as *rT*, while the correlation for the non-target speech is denoted as *rM*. The difference between *rT* and *rM*, termed *rD*, was used to quantify selective neural speech entrainment, with *rD* > 0 indicating a larger neural tracking of the target talker relative to the masker talker and hence, a successfully decoded trial.

III. Results

During the attentive listening task, participants were instructed to focus on the target speech and press a buzzer when repeated words appeared in the target speech at random time intervals (mean interval = 10s). Word detection accuracy (DA) was computed and fitted to a psychometric curve to reflect average speech perception, providing a better estimate of speech intelligibility for long, continuous stimuli by considering participants' real-time status, as shown in Figure 2.

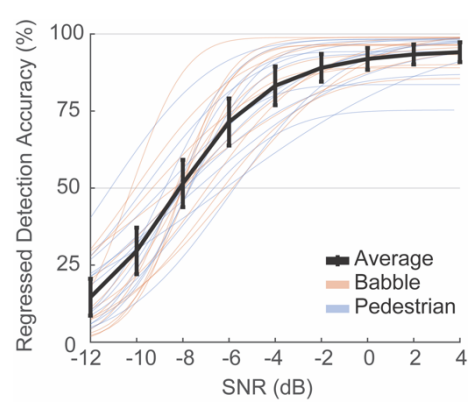

Fig. 2. Individual psychomtetric curves for two types of background noises (red and blue) and the averaged psychometric curve (black).

These curves represent individual responses to two different background noise conditions (red and blue lines). The figure also shows the averaged psychometric curve (black line).

Figure 2 highlights how DA increases as the SNR improves until it reaches a ceiling. Next, we examined how the neural tracking of target speech changes with varying DA.

A. Parameters influencing AAD performance

AAD accuracy is defined as the percentage of trials where the target speech is successfully identified. In other words, if the correlation between neural responses and the target speech envelope (*rT*) is greater than that of the masker speech (*rM*): $rD = rT - rM > 0$. As hypothesized, both decoding duration and DA are positively correlated with single-trial *rD* values and AAD accuracy, as illustrated in Figure 3A and 3B. The statistical significance of comparing *rD* with zero is demonstrated in Figure 3C. Figure 3 shows that as the intelligibility of the target speech decreases, the required decoding duration of the trial for a successful neural decoding of the target speech also increases. However, neural decoding is still possible even when the target speech is barely detectable as low as 25%, which is when only one out of each four words are heard.

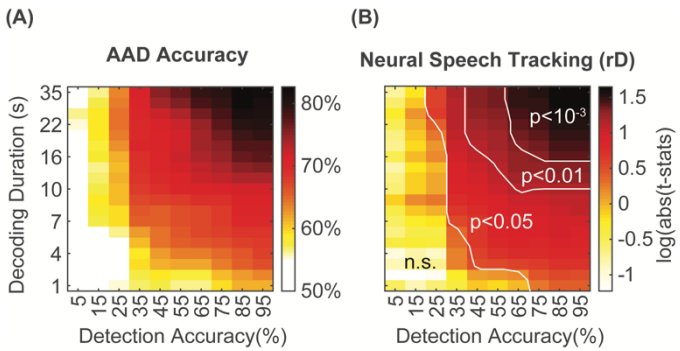

Fig. 3. (A). Bootstrap AAD accurary. (B). Boostrap log(abs(t-tstat)) by comparing *rD* with zero. Bonferroni corrected p-val by comparing *rD* with zero using t-test. 300 rounds bootstrap with 30 samples each time was applied to alleviate samples imbalance.

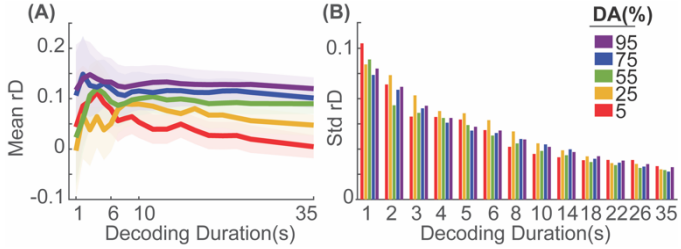

Fig. 4. (A) Bootstrap average *rD* for different DA groups, across different decoding durations. (B) Boostrap standard deviation of rD for different DA groups. 300 rounds bootstrap with 30 samples each time was applied to alleviate samples imbalance.

We conducted further analysis to isolate the effects of decoding duration and DA on *rD*. As shown in Figure 4A, the mean of *rD* converges and remains constant after about 6 seconds, where more intelligible trials (higher DA) result in higher values of *rD*. Conversely, the standard deviation of *rD* is unaffected by DA and is solely determined by decoding duration as shown in Figure 4B. In conclusion, increasing DA improves *rD*, while increasing decoding duration reduces the statistical detectability of the single-trial AAD outcomes.

B. Simulation of AAD's benefit

Since we have shown that AAD can be performed even in low intelligibility conditions, this suggests that AAD could be used to improve SNR in such scenarios, potentially leading to enhanced intelligibility. To explore this, we simulated the impact of AAD by comparing AAD-on and AAD-off conditions, as shown with red and black lines in Figure 5 accordingly. In AAD-on condition, we do the following:

- When the predicted target speech is louder than the predicted non-target speech, the overall loudness is adjusted to maintain 65 dB for the target, ensuring a consistent listening experience. No further amplification or noise reduction is applied to alter the relative amplitude of the target and non-target speech, allowing for smooth attentional switches between speakers and minimizes detection errors.
- When the predicted target speech is quieter than the predicted non-target speech, the target speech is amplified to 65 dB, and the non-target speech is reduced to achieve a 0 dB SNR for the target. Specifically, under that SNR, the non-target is still fully detectable.

In other words, the AAD-on condition operates on the left side of the psychometric curve and lifts up the detection accuracy by increasing the SNR of the target speech. This simulation highlights the potential for AAD to significantly enhance speech perception in challenging listening conditions.

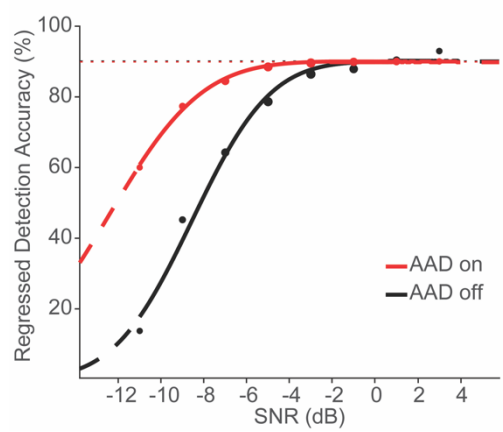

Fig. 5. Simulation of the modulation effect on psychometric curves by implementing AAD (AAD-on, red line), compared to baseline (AAD-off, black line).

IV. Discussion

We investigated how the brain tracks attended speech in multi-talker environments with varying levels of background noise. Using a repeated-word detection task and EEG recordings, we demonstrated that selective neural tracking of the attended speaker remains robust even when intelligibility is significantly degraded. However, achieving the same auditory

attention decoding (AAD) accuracy requires a longer decoding duration under these challenging conditions. Our findings provide new insights into the resilience of neural tracking in adverse environments and its potential implications for real-world AAD applications.

Speech perception was measured using a repeated-word detection task, providing a dynamic and real-time proxy for perceptual accuracy that accounts for both perceptual and cognitive factors [22], [23]. Unlike traditional speech intelligibility measures, this approach captures a broader spectrum of influences on perception. Importantly, our results challenge the assumption that high intelligibility is necessary for effective neural decoding [24]. We showed that selective neural tracking persists even when intelligibility is significantly degraded, extending the potential of AAD to low-intelligibility scenarios. For example, neural tracking remained robust when only one out of every four words was perceived (DA = 25%). This contrasts with previous research suggesting that intelligibility is critical for successful neural tracking [24].

Our findings also revealed a strong positive correlation between neural speech tracking and speech perception, aligning with prior studies [9], [12], [24], [25]. However, a key contribution of this study is the demonstration that neural tracking can be sustained even in environments with severely degraded intelligibility. This opens up new possibilities for applying AAD in real-world scenarios, where speech is often masked by noise. Additionally, we simulated the effect of AAD on speech perception and found that implementing AAD can significantly enhance perception, particularly in low-SNR conditions. These findings underscore the potential of AAD to optimize listening experiences in challenging environments. However, further research is needed to validate these findings in real-time, closed-loop systems and to explore the impact of cognitive factors such as attention and effort on decoding performance.

In future work, we plan to extend this research to more complex auditory scenes, where attended conversations involve turn-taking speakers who may also move in space rather than remain stationary [26]. We also aim to test these findings in individuals with hearing impairments to ensure the applicability of AAD in populations that stand to benefit most from enhanced speech perception technologies. Addressing these challenges will pave the way for developing adaptive, brain-informed technologies that enhance listening experiences in complex auditory environments.

References

[1] J. R. Kerlin, A. J. Shahin, and L. M. Miller, "Attentional Gain Control of Ongoing Cortical Speech Representations in a 'Cocktail Party,'" *J. Neurosci.*, vol. 30, no. 2, pp. 620–628, Jan. 2010.

[2] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, May 2012.

[3] M. Koskinen, J. Viinikanoja, M. Kurimo, A. Klami, S. Kaski, and R. Hari, "Identifying fragments of natural speech from the listener's MEG signals," *Hum. Brain Mapp.*, vol. 34, no. 6, pp. 1477–1489, 2013.

[4] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Phoneme representation and classification in primary auditory cortex," *J. Acoust. Soc. Am.*, vol. 123, no. 2, pp. 899–909, Feb. 2008.

[5] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Mechanisms of noise robust representation of speech in primary auditory cortex," *Proc. Natl. Acad. Sci.*, vol. 111, no. 18, pp. 6792–6797, May 2014.

[6] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proc. Natl. Acad. Sci.*, vol. 109, no. 29, pp. 11854–11859, Jul. 2012.

[7] J. A. O'Sullivan *et al.*, "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cereb. Cortex N. Y. N 1991*, vol. 25, no. 7, pp. 1697–1706, Jul. 2015,.

[8] D. Lesenfants, J. Vanthornhout, E. Verschueren, L. Decruy, and T. Francart, "Predicting individual speech intelligibility from the cortical tracking of acoustic- and phonetic-level speech representations," *Hear. Res.*, vol. 380, pp. 1–9, Sep. 2019.

[9] J. Vanthornhout, L. Decruy, J. Wouters, J. Z. Simon, and T. Francart, "Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope," *J. Assoc. Res. Otolaryngol. JARO*, vol. 19, no. 2, pp. 181–191, Apr. 2018.

[10] L. Decruy, J. Vanthornhout, and T. Francart, "Evidence for enhanced neural tracking of the speech envelope underlying age-related speech-in-noise difficulties," *J. Neurophysiol.*, vol. 122, no. 2, pp. 601–615, Aug. 2019.

[11] O. Etard and T. Reichenbach, "Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise," *J. Neurosci.*, vol. 39, no. 29, pp. 5750–5759, Jul. 2019.

[12] I. Iotzov and L. C. Parra, "EEG can predict speech intelligibility," *J. Neural Eng.*, vol. 16, no. 3, p. 036008, Jun. 2019.

[13] W. K. Kirchner, "Age differences in short-term retention of rapidly changing information," *J. Exp. Psychol.*, vol. 55, no. 4, pp. 352–358, Apr. 1958.

[14] S. Dähne *et al.*, "Multivariate Machine Learning Methods for Fusing Multimodal Functional Neuroimaging Data," *Proc. IEEE*, vol. 103, no. 9, pp. 1507–1530, Sep. 2015.

[15] A. de Cheveigné, D. D. E. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, May 2018.

[16] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[17] S. S. Kang, T. J. Lano, and S. R. Sponheim, "Distortions in EEG interregional phase synchrony by spherical spline interpolation: causes and remedies," *Neuropsychiatr. Electrophysiol.*, vol. 1, no. 1, p. 9, Aug. 2015.

[18] F. Perrin, J. Pernier, O. Bertrand, and J. F. Echallier, "Spherical splines for scalp potential and current density mapping," *Electroencephalogr. Clin. Neurophysiol.*, vol. 72, no. 2, pp. 184–187, Feb. 1989.

[19] R. L. Horwitz-Martin, T. F. Quatieri, E. Godoy, and J. R. Williamson, "A vocal modulation model with application to predicting depression severity," in *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, Jun. 2016, pp. 247–253.

[20] F. A. Wichmann and N. J. Hill, "The psychometric function: I. Fitting, sampling, and goodness of fit," *Percept. Psychophys.*, vol. 63, no. 8, pp. 1293–1313, Nov. 2001.

[21] F. A. Wichmann and N. J. Hill, "The psychometric function: II. Bootstrap-based confidence intervals and sampling," *Percept. Psychophys.*, vol. 63, no. 8, pp. 1314–1329, Nov. 2001.

[22] E. M. J. Devocht, A. M. L. Janssen, J. Chalupper, R. J. Stokroos, and E. L. J. George, "The Benefits of Bimodal Aiding on Extended Dimensions of Speech Perception: Intelligibility, Listening Effort, and Sound Quality," *Trends Hear.*, vol. 21, p. 2331216517727900, Sep. 2017.

[23] S. Gatehouse and W. Noble, "The Speech, Spatial and Qualities of Hearing Scale (SSQ)," *Int. J. Audiol.*, vol. 43, no. 2, pp. 85–99, Feb. 2004.

[24] N. Ding and J. Z. Simon, "Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical Representation of Speech," *J. Neurosci.*, vol. 33, no. 13, pp. 5728–5735, Mar. 2013.

[25] L. Riecke, E. Formisano, B. Sorger, D. Başkent, and E. Gaudrain, "Neural Entrainment to Speech Modulates Speech Intelligibility," *Curr. Biol. CB*, vol. 28, no. 2, pp. 161-169.e5, Jan. 2018.

[26] V. Choudhari *et al.*, "Brain-Controlled Augmented Hearing for Spatially Moving Conversations in Multi-Talker Environments," *Advanced Science.*