Auditory and Vestibular Research

# Investigating Central Compensation For Voice Onset Time In Noise Using Deep Learning

**Amirhossein Sameti[1], Nematollah Rouhbakhsh[1*], Amir Homayoun Jafari [2,3], Zahra Shirzhiyan[4]**

[1] Audiology Department, School of Rehabilitation, Tehran, University of Medical Sciences, Tehran, Iran.

[2] Department of Medical Physics and Biomedical Engineering, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran

[3] Research Center for Biomedical Technologies and Robotics, Tehran University of Medical Sciences, Tehran, Iran.

[4] Computational Neuroscience, Institute of Medical Technology, Brandenburg University of Technology Cottbus-Senftenberg, Cottbus, Germany.

*Corresponding author: Audiology Department, School of Rehabilitation, Tehran University of Medical Sciences, Tehran, Iran. rohbakhn@tums.ac.ir



## Highlights

- EEG-derived $r_{cc}$ and classification metrics index central auditory compensation
- Higher $r_{cc}$ and CNN classification in noise predict better speech-in-noise performance
- The Fronto-central cortex shows the most noise-invariant speech representations

## Abstract

**Background and aim:** The brain's ability to resolve rapid temporal cues such as voice-onset time (VOT) is essential for speech perception in challenging listening environments. We tested whether central auditory compensation for VOT in noise is reflected in the fidelity of cortical auditory evoked potentials (CAEPs) using a neural-network classifier and a cross-condition similarity metric.

**Methods:** Electroencephalography (EEG) was recorded from 22 normal-hearing adults in response to /ka/ and /ga/ syllables with varying VOTs, presented in quiet and noise (+7 dB Signal-to-noise ratio). We measured CAEPs' peak amplitude (N1-P2), employed a convolutional neural network (CNN) to classify CAEPs by syllable identity, and computed a cross-condition correlation ($r_{cc}$) to quantify the similarity between responses in quiet and noise.

**Results:** Background noise significantly reduced N1-P2 amplitude, behavioral performance, and CNN classification accuracy, confirming the degradation of phoneme-specific neural representations. Critically, inter-subject variability in behavioral speech in noise performance was significantly correlated by both $r_{cc}$ (*r=0.443, p=0.02**) and CNN accuracy in noise (*r=0.492, p=0.01**). Individuals with higher behavioral speech-in-noise (SIN) scores exhibited CAEPs in noise that were more similar to their clean-speech responses (higher $r_{cc}$) and more discriminable by CNN. Scalp topography displayed the highest $r_{cc}$ values over fronto-central regions, with the strongest correlation between $r_{cc}$ and SIN performance.

**Conclusion:** The convergence of our findings demonstrates that successful SIN perception relies on the brain's capacity to maintain a stable, noise-invariant cortical representation of speech, particularly in fronto-central auditory regions. These EEG-derived metrics may serve as a research tool for future clinical investigations.

**Keywords**: speech-in-noise perception, voice-onset time, cortical auditory evoked potentials, convolutional neural network, electroencephalography, temporal processing.

## Introduction

Auditory temporal processing ability plays a central role in speech, language, and reading skills [1]. Temporal processing refers to the auditory system's ability to encode and analyze temporal (timing) cues in the speech signal, like voice-onset time (VOT), stimulus duration, and amplitude modulation [2-4]. Disorders of temporal processing are thought to underlie impaired discrimination abilities and reduced speech-in-noise intelligibility. Therefore, intact processing of temporal speech cues is required for understanding speech in noise [5,6]. Asynchronous or jittered timing is a factor underlying reduced coding of dynamic speech features and speech in noise [7].

A key temporal cue for speech perception is VOT, defined as the interval between the stop consonant's release and the start of vocal fold vibration. Rapid temporal changes like VOT allow listeners to discriminate similar consonants (for example, /da/ vs. /ta/, /ba/ vs. /pa/, /ga/ vs. /ka/) [8].

In general, VOT for voiced syllables is relatively brief (e.g., <45 milliseconds in /ga/), whereas voiceless syllables have longer VOT duration (e.g., >45 milliseconds in /ka/). VOTs that fall on categorical boundaries are classified less consistently as /k/ or /g/ than VOTs at the extremes of the continuum [9]. Accurate identification of VOT requires precisely time-locked neural responses to voicing onset and rapid spectral–temporal processes in the auditory cortex [10-12].

Importantly, masking noise weakens and delays neural speech encoding. Degraded cortical representation of fine temporal cues like VOT directly impairs the brain's ability to discriminate phonemes in noisy conditions [13].

Cortical auditory evoked potentials (CAEPs) offer an objective, high-temporal-resolution measure to evaluate cortical speech processes. Robust CAEPs rely on precise neural synchronization to stimulus onset. Preservation of such time-locked responses is essential for speech perception in both quiet and background noise [14]. Thus, examining CAEPs under clean and competing-noise conditions can yield valuable neurophysiological insight into speech perception difficulties.

Adult late-latency CAEPs consist of sequential negative and positive deflections identified by the N1–P2 complex, typically occurring about 100–200 ms post-stimulus. One advantage of recording the N1–P2 complex is that it does not require directed participants' attention and therefore serves as an index of automatic cortical processing of temporal characteristics on the auditory sensory level [15]. CAEPs have been widely used to objectively assess neural encoding of VOT at the cortical level in both healthy listeners and clinical populations.

The N1component, a negative peak at ≈100 ms, is sensitive to temporal changes and serves as a reliable indicator of neural speech timing at the cortical level [16,17]. P2, a positive peak at ≈200 ms, reflects sound content properties such as acoustic or phonetic structures [18]. The N1–P2 complex, as commonly used to measure the neural mechanisms underlying VOT, has been shown to change with VOT and to correlate with behavioral identification of syllables varying in VOT [19,20].

However, some inconsistencies remain in previous studies. Some of them report delayed N1s with longer VOTs [21,22], whereas others report larger N1 amplitudes for short VOTs without any accompanying changes in latency

[23,24]. Similar discrepancies appear when stimuli are presented in background noise. For example, unlike quiet, there is no significant difference between N1 amplitudes in response to /du-tu/ syllables with different VOTs when presented in background noise [23]. Some of these inconsistencies may arise from using simple traditional features when analyzing evoked potentials. But not all relevant neural characteristics are limited to response amplitude or latency.

Modern approaches, such as deep learning, allow for decoding stimulus-related features from electrophysiological signals via learned representations that outperform conventional amplitude and latency measures. Such decoding provides a more sensitive test of discriminability among neural responses, revealing how effectively they convey information about phonetic classes beyond traditional evoked potentials comparison [25].

In the current study, late-latency CAEPs were recorded in response to syllables with VOTs chosen from two adjacent perceptual categories as well as the boundary between them. Syllables were presented in two listening conditions: quiet and noise. Recorded responses were then analyzed using a convolutional neural network (CNN) to better investigate cortical encoding of VOT with its behavioral correlates under both quiet and noise conditions. This approach also allowed examination of how noise affects the cortical processing of VOT relative to the perceptual boundary.

In parallel with the classification analysis, we sought a physiologic measure of how effectively the central auditory system resists background noise. To this end, we employed a cross-condition correlation ($r_{cc}$) metric, which directly quantifies the similarity between neural responses recorded in quiet and noise. We hypothesized that higher $r_{cc}$ values reflect a neural capacity to suppress the masking effects of noise, thereby preserving a better identification performance in background noise.

## Method

### Participants and electroencephalography acquisition

Twenty-two male listeners aged 19 to 33 were included in the study. All the subjects were screened to ensure pure-tone thresholds of ≤20 dB hearing level (HL) at 0.25, 0.5, 1, 2, and 4 kHz. The experiment was conducted at the Iran National Brain Mapping Laboratory, Tehran. Electroencephalography (EEG) signals were recorded at a sampling rate of 1200 Hz using a 32-channel stable system (g.tec medical engineering GmbH, Austria) comprising a head cap, an amplifier, and a Windows computer. The signal was referenced to linked earlobes and the ground electrode was placed at Fpz.

Thirty-two active electrodes were used, with linked earlobes (connected by a jumper) serving as the reference electrode. Each electrode cup was filled with conductive gel (Onestep Cleargel). Signals were recorded with a sampling frequency of 1200 Hz and bandpass-filtered online from 0.5 to 500 Hz. Data were then downsampled to 300 Hz, and then a 2-Hz high-pass finite impulse response (FIR) filter (12 dB/octave slope) was applied. EEG Signals were epoched from −100 ms pre-stimulus to 1000 ms post-stimulus onset. Trials containing activity exceeding ± 75 µv were rejected as artifacts. The remaining trials were used to attain CAEPs.

### Stimuli and presentation paradigms

Three synthetic /ga/–/ka/ consonant–vowel (CV) syllables were generated in Praat [26] with VOTs of 35, 45, 55 ms and were presented in a free field condition using PejvakAva loudspeaker system. The event trigger pulses were simultaneously delivered to the EEG amplifier through a parallel port. Sound level was calibrated using a Sound Level Meter (Norsonic AS, Norway). The temporal waveforms and spectrograms of the stimuli are shown in Figure 1.

The syllables with VOTs of 35-ms and 55-ms fell well within the perceptual domain of /ga/ and /ka/, respectively, whereas the syllable with the VOT of 45-ms lies on the boundary between them. Syllables were presented in two listening conditions: quiet and noise. The duration of stimuli was 195 ms. Syllables were presented every 1.1 seconds. Stimuli were presented at 70 dB SPL. Background noise was continuous white noise presented at +7 dB signal-to-noise ratio (SNR). A moderate SNR was employed, as lower SNR levels might necessitate invasive recording methods, such as electrocorticography, to extract speech-related features [27]. White noise was selected for this study due to its flat spectrogram and lack of temporal fluctuations, which prevents the evocation of noise-induced onset responses that could interfere with auditory late latency responses.

Each of the syllables was presented 160 times in each listening condition. Because there were three syllables and two sound conditions, a total of 960 trials were presented per subject (2 × 3 × 160); additionally, 2% of the trials were rejected due to artifacts. The listening condition (quiet vs. noise) remained constant within each recording session, whereas syllable order was randomized.

After EEG data collection, the stimuli were presented again to measure subjects' identification scores by asking them to choose the correct presented syllable on a close-set questionnaire. EEG and behavioral data were not recorded simultaneously to avoid motor-related neural activity (e.g., from manual responses) contaminating CAEPs or topographies.

**Cortical auditory evoked potential peak analysis**

We examined the N1-P2 complex that was most prominent in the CAEPs. Not all electrodes or subjects showed both peaks; Thus, in the results, responses of the electrodes that consistently exhibited the N1-P2 complex in the majority of the participants are presented. After the mean peak latencies had been determined, a window of 220 ms was searched to identify the peaks' maximum or minimum values.

**Cortical auditory evoked potentials classification**

Signal classification evaluates how CAEPs of different syllables were *consistently different* from one another. A "leave-one-subject-out-cross-validation" approach was used to provide the training and test samples for the classifier. For each fold, one subject was held out as the test subject, and the remaining subjects were set as the training set. When the model was trained, it was evaluated on the held-out test subject.

Combined trials were used to produce training or test samples. For each fold, to build the training and test sets, 20 trials were randomly sampled with replacement from each subject's recorded trials and averaged to produce a single CAEP; this procedure was repeated about 70 times per participant, yielding approximately 70 CAEPs per participant and approximately 4,200 CAEPs in total for the training set in each fold.

A CNN was then used to classify CAEPs elicited by different syllables. The first layer applied a 2-D convolutional filter of size (1,64); the filter length was chosen to be approximately one-quarter of the sampling rate (here, 300 Hz), to capture frequency information at approximately 4 Hz and above. A subsequent convolutional layer of size (channels,1) was used to learn spatial filters. The output is then 1D feature maps of size (1,330) containing combined learned temporal and spatial features. Both convolutional layers were kept linear because no performance gains were observed with nonlinear activations. Then batch normalization was applied along the feature-map dimension, followed by an Exponential Linear Unit (ELU) activation.

To help regularize, the dropout technique was used with a probability of 0.25 to help prevent overfitting. An average pooling layer of size (1,2) was applied for temporal downsampling. Convolutional layers were further regularized using a maximum-norm constraint of 2 on the weights ($\|w\|_2 < 2$). A subsequent convolutional layer

with kernel length (1,32) learned a summary over approximately 215 ms of EEG activity (at an effective sampling rate of 150 Hz), containing both the extracted temporal and spatial feature maps. Then an average pooling of size (1,8) was used for dimension reduction. In the classification unit, the extracted features were passed to a Softmax layer with 3 units corresponding to three classes. Model architecture is detailed in Table 1. The model was trained using the Adam optimization algorithm over 100 epochs with a batch size of 32 and a learning rate of 0.001. Categorical cross-entropy was employed as the loss function, given the multi-class nature of the classification task.

The derived accuracy indicates, for a given sound condition, how consistently the CAEPs elicited by different syllables were distinguishable. By comparing accuracies obtained in the noise and quiet conditions, the effect of noise can be assessed. Furthermore, we evaluated the correlation between $\hat{p}$ and behavioral performance across subjects to determine *whether high speech intelligibility was achieved together with distinguishable CAEPs in the brain signals.*

**Electroencephalography cross-condition correlations**

Because three syllables were used, each listening condition elicited three CAEPs, each averaged over 160 trials. For a given electrode, the CAEPs for the three syllables were concatenated in time to compute a single $r_{cc}$ value. This approach allows the resulting $r_{cc}$ be accounted for amplitude differences across CAEPs evoked by different syllables and is therefore preferred to averaging separate correlations.

To obtain $r_{cc}$, correlations were computed between the CAEPs recorded in the control (quiet) condition and those recorded in the noise. The derived $r_{cc}$ indicates *waveform similarity* between clean and noisy CAEPs. These $r_{cc}$ values were further compared with behavior performance to assess *whether high speech intelligibility is associated with a greater CAEP similarity between clean-speech and background noise conditions.*

**Results**

**Cortical auditory evoked potentials and speech-related features**

Figure 2a shows examples of CAEPs for the silence condition, and Figure 2b shows CAEPs recorded in the noise condition, each averaged over 160 trials, obtained with electrode FC2.

CAEPs evoked by speech in noise had reduced amplitude and increased latency compared to silent condition. The most consistent features were N1 and P2 occurring at approximately 140 ms and 200 ms post-stimulus, as marked in Figure 2. Since not all peaks are always present at every electrode or in every subject, Figure 2 presents waveforms from the FC2 electrode, which consistently exhibited the N1-P2 complex in the majority of participants. As presented in Figure 2, although recorded responses in the noise condition show a marked reduction in N1–P2 amplitude compared to the silent condition, most subjects exhibited largely similar CAEPs across the three syllables in both the quiet and noise conditions.

To quantify the effect of noise on CAEPs, the absolute N1-P2 amplitudes were extracted. Figure 3 shows the results for representative electrodes. The *x*-axis is always the peak height in the control (clean-speech) condition, and the *y*-axis is the peak height for syllables presented in noise. Symbols below the diagonal line would indicate that the peak became smaller when noise was added.

Each symbol denotes the mean value for a single syllable averaged over 160 trials, and hence, there are three symbols per condition for each subject, with a total of twenty-two subjects. As presented in Figure 3, in most of the auditory-related channels, the N1-P2 amplitudes were significantly reduced when noise was added to the background.

**Effect of noise on classification accuracy**

Recall that high accuracy indicates that CAEPs elicited by the three syllables were consistently different and, therefore, discriminable. Chance performance for three classes is 33.3%. Figure 4 shows model accuracy per participant for two silence and noise conditions, where asterisks' sizes indicate false discovery rate (FDR) corrected significance level from the one-sided binomial test of model performance compared to chance level for each participant under each listening condition. Prior to any analysis of model accuracy or behavioral performance, the normality of the data was confirmed using a one-sample Kolmogorov–Smirnov test. To compare the discriminability of CAEPs between conditions, since measurements were paired within subjects, the paired-sample t-test was used on subject-level data. The results show that the model performance is significantly higher in the silence condition compared to the noise condition (*paired* $t$=3.452, $p$=0.002, *d=0.736*), indicating that CAEPs were significantly more discriminable in the silence condition versus the noise condition. This result is also reflected in behavioral performance, with significantly lower performance in the silent compared to the background noise condition across subjects (*paired t=5.45, p<0.001, d= 1.161*).

A further considerable finding was that adding noise increased the CAEPs' confusion between the border (VOT= 45 ms) and non-border syllables significantly more than it increased confusion between the two non-border syllables (*paired t=2.335*, *p=0.015*, *d=0.510*).

**Noise and cross-condition correlations**

Figures 3 and 4 show that prominent CAEP components (e.g., N1-P2 complex) and phonem-related neural characteristics are significantly affected by background noise. Additional CAEPs features can be altered as well, which can be assessed by computing $r_{cc}$ between CAEPs recorded in the control (i.e., clean speech) and other listening condition. Figure 5a shows the $r_{cc}$ scalp topography of CAEPs averaged over subjects for all recording channels. Warm colors indicate high correlation values. The highest $r_{CC}$ values were observed over fronto-central regions, including Fz, FC1, FC2, Cz channels associated with N1–P2 sources.

Figure 5b shows the scalp topography of the correlation between $r_{cc}$ and the behavioral responses. Consistent with the $r_{cc}$ scalp map, the highest correlations were again achieved by fronto-central regions, indicating electrodes in these regions were most consistent in explaining human behavior using $r_{cc}$.

**Electroencephalography physiology and behavioral performance**

Adding background noise increased the difficulty of behavioral syllable identification, as evidenced by a decrease in the average correct identification score from 77.39% in the silent condition to 62.30% in the noise condition. Next, we examined whether the behavioral performance agreed with CAEP measures (classification accuracy and/or *r* cc values) in terms of cross-subject variations.

To this end, the CAEP measures related to noise condition, such as $r_{cc}$ or classification accuracy in noise, were examined with behavioral performance in noise, and the classification accuracy in silence was examined with behavioral accuracy in the silence condition.

Figure 6a and 6b show scatterplots with the linear fit in terms of the correlation between model accuracy and behavioral performance in silence and noise, respectively. There was a significant positive correlation between model accuracy and behavioral performance in silent ($r = 0.562$, $p=0.003$**, $r^2=0.316$) and in noise ($r = 0.492$, $p=0.01$*, $r^2=0.242$), indicating that subjects with better behavioral discrimination also had more discriminable CAEPs to syllables.

Figure 6c shows the scatterplot of $r_{cc}$ values achieved by the FC2 electrode and behavioral performance in noise. Recall that $r_{cc}$ measures the similarity between CAEPs recorded in clean speech and noise conditions. There was a significant positive correlation between $r_{cc}$ and behavioral performance in noise ($r = 0.443$, $p=0.02$*, $r^2=0.196$), indicating subjects with higher identification scores also exhibited greater similarity of CAEPs between clean and noisy speech by suppressing brain responses to noise.

## Discussion

The present study investigated the cortical correlates of speech-in-noise (SIN) perception by examining how background noise affects the neural encoding of VOT and, critically, how individual differences in neural metrics predict behavioral performance. Our findings confirm that noise degrades the cortical representation of VOT, as evidenced by N1-P2attenuated amplitudes and reduced deep learning-based classification accuracy of syllables varying in VOT. More importantly, we demonstrate that successful SIN perception is linked to two key neural properties: the distinctiveness of phoneme-related cortical patterns (as measured by CNN classification) and the fidelity or noise-invariance of the overall cortical response (as measured by $r_{cc}$). Although practical constraints prevented a larger sample size, the current sample provided reasonably good statistical power given the reported effect sizes. Specifically, the observed power reached 95.5% for the correlation analysis between model accuracy and behavioral performance, and 88.4% for the paired t-test comparing model performance under silent and noise conditions—the two primary objectives of this study.

### Cortical degradation of temporal cues in noise

Consistent with prior literature [23], we observed a significant reduction in the amplitude of the N1-P2 complex when syllables were presented in noise. This noise-induced suppression reflects a degradation of the synchronized cortical activity necessary for processing rapid acoustic transients. The concomitant decline in CNN classification accuracy in the noise condition provides a more nuanced confirmation of this degradation.

While the N1-P2 complex represents a gross neural response, the CNN's performance is contingent on the fine-grained, distributed pattern of activity that distinguishes one phoneme from another. The significant drop in classification accuracy from silence to noise indicates that these critical, phoneme-identifying features within the CAEPs are being masked or altered, implying the brain's impaired ability to build distinct neural representations of characteristic features— as previously shown for formant structure encoding [27]. This finding also aligns with and extends previous work reporting no significant differences in N100 responses to syllables with different VOTs presented in background noise [23].

### Fronto-central cortex as a hub for robust speech encoding

Our topographical analyses consistently highlighted the fronto-central scalp region (electrodes Fz, FC1, FC2, Cz) as a critical hub. This area not only exhibited the highest $r_{cc}$ values—indicating the most noise-invariant responses—but also showed the strongest correlation between $r_{cc}$ and SIN performance. Frontal cortical areas are known to overlie activity that can directly influence the selectivity of auditory cortex neurons, enhancing responses to target stimuli while suppressing responses to noise [28,29]. This active modulation occurs through top-dwon pathways from frontal regions to the auditory cortex, indicating the role of cognitive control in listening in noise.

Our results suggest that these regions are not merely passive recipients of degraded auditory input but are actively involved in generating a stable, noise-resistant representation of speech. The reliability of syllable-specific

waveforms in this area and its established role in speech perception in noise [28,29] support its identification here as a primary locus for the central compensation mechanisms we measured.

**Linking neural fidelity to behavioral perception: the role of cross-condition correlation and classification**

The core novel finding of this study is the significant correlation between behavioral speech in noise performance and our two EEG-derived metrics. The positive correlation with CNN accuracy indicates that individuals whose brains generate more distinct cortical patterns for different phonemes are better able to identify those phonemes behaviorally. It implies that subjects' ability to distinguish syllables with different speech characteristics (e.g., VOT, formant) primarily reflects cortical processing of the corresponding sensory-level features. Furthermore, the difference in adverse effects of noise near the perceptual boundary indicates that these effects are modulated by higher-order perceptual processes, revealing a mutual connection between cortical representation of the sensory inputs and perceptually embedded schemes. This finding aligns with recent studies demonstrating that the sensory processes of formant structures (as measured by frequency-following responses) are not only sensitive to acoustic features but are also modulated by listeners' perceptual categorization, especially near category boundaries [30].

The correlation with $r_{cc}$, however, provides a complementary and perhaps more fundamental insight. $r_{cc}$ measures the global similarity between the clean and noisy CAEP waveforms, essentially quantifying how well the brain ignores the noise to produce a response akin to that evoked by clean speech.

Consistent with previous studies [27], the significant positive correlation between $r_{cc}$ and behavior suggests that individuals with a superior innate or learned capacity for subcortical or early cortical noise suppression would likely make fewer identification errors by providing a cleaner signal to higher-order processing centers. This finding provides a direct physiological answer to a key question in auditory neuroscience: robust speech in noise perception is partly achieved by neural mechanisms that preserve the integrity of the speech-evoked cortical response amidst interfering noise, a phenomenon we term cortical noise suppression.

**Synthesis and implications**

Taken together, our results support a two-stage model for successful speech-in-noise perception in the cortex: first, a noise suppression mechanism (indexed by $r_{cc}$) that preserves the overall fidelity of the auditory object representation; and second, a pattern discrimination mechanism (indexed by CNN accuracy) that allows for the extraction of critical phonemic features from this stabilized representation. The failure of either process leads to behavioral deficits.

From a clinical perspective, the $r_{cc}$ metric, in particular, offers a promising, objective tool for assessing central auditory processing disorders. It moves beyond simply measuring degradation (e.g., reduced N1 amplitude) to quantifying a positive, compensatory function. This could be invaluable for diagnosing populations known to have disproportionate speech in noise difficulties, such as individuals with hearing loss, auditory processing disorder, or dyslexia, and for evaluating the efficacy of interventions like hearing aids or auditory training.

**Conclusion**

In conclusion, by leveraging both traditional and advanced computational analyses of cortical activity, we have demonstrated that robust speech perception in noise is predicated on the brain's ability to maintain stable and distinct neural representations of speech sounds. We introduce the $r_{cc}$ as a robust neural correlate of central noise suppression. Our findings underscore that the challenge of hearing in noise is not merely a problem of peripheral

masking but is fundamentally a challenge of central neural fidelity, which can be directly measured and linked to perceptual ability.

## Ethical Considerations

### Compliance with ethical guidelines

This study has ethical approval from Research Ethics Committees of Tehran University of Medical science (Approval ID. IR.TUMS.FNM.REC.1402.092).


### Funding

This study is funded by Tehran University of Medical science (Grant No. 1402-4-103-66962).


### Authors' contributions

AS: Study design, acquisition of data, interpretation of the results, EEG Analyzing, statistical analysis, Writing – original draft; NR: Supervision, Funding acquisition, Methodology, Study design, interpretation of the results, Writing – review & editing; AJ: Study design, interpretation of the results, Writing – review & editing; ZH: Study design, interpretation of the results, EEG Analyzing, Writing – review & editing.

### Conflict of interest

The authors declare that they have no competing interest.


### Acknowledgments

This article is derived from the doctoral dissertation of the first author, Amirhossein Sameti (Registration No. 40011303001). The authors gratefully acknowledge the subjects who participated in this study for their patience and cooperation.




## References


1. Han JH, Zhang F, Kadis DS, Houston LM, Samy RN, Smith ML, et al. Auditory cortical activity to different voice onset times in cochlear implant users. Clin Neurophysiol. 2016;127(2):1603-1617. [DOI:10.1016/j.clinph.2015.10.049]
2. Al-Meqbel A, McMahon C. Cortical auditory temporal processing abilities in elderly listeners. 2015;24(2):80-91.
3. Fox NP, Leonard M, Sjerps MJ, Chang EF. Transformation of a temporal speech cue to a spatial neural code in human auditory cortex. Elife. 2020;9:e53051. [DOI:10.7554/eLife.53051]
4. Parida S, Yurasits K, Cancel VE, Zink ME, Mitchell C, Ziliak MC, et al. Rapid and objective assessment of auditory temporal processing using dynamic amplitude-modulated stimuli. Commun Biol. 2024;7(1):1517. [DOI:10.1038/s42003-024-07187-1]
5. Tallal P, Miller SL, Bedi G, Byma G, Wang X, Nagarajan SS, et al. Language comprehension in language-learning impaired children improved with acoustically modified speech. Science. 1996;271(5245):81-4. [DOI:10.1126/science.271.5245.81]
6. Johannesen PT, Pérez-González P, Kalluri S, Blanco JL, Lopez-Poveda EA. The Influence of Cochlear Mechanical Dysfunction, Temporal Processing Deficits, and Age on the Intelligibility of Audible Speech in Noise for Hearing-Impaired Listeners. Trends Hear. 2016;20:2331216516641055. [DOI:10.1177/2331216516641055]
7. White-Schwoch T, Nicol T, Warrier CM, Abrams DA, Kraus N. Individual Differences in Human Auditory Processing: Insights From Single-Trial Auditory Midbrain Activity in an Animal Model. Cereb Cortex. 2017;27(11):5095-115. [DOI:10.1093/cercor/bhw293]
8. Abramson AS, Whalen DH. Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. J Phon. 2017;63:75-86. [DOI:10.1016/j.wocn.2017.05.002]
9. Sharma A, Marsh CM, Dorman MF. Relationship between N1 evoked potential morphology and the perception of voicing. J Acoust Soc Am. 2000;108(6):3030-5. [DOI:10.1121/1.1320474]

10. Tallal P. Language disabilities in children: perceptual correlates. Int J Pediatr Otorhinolaryngol. 1981;3(1):1-13. [DOI:10.1016/0165-5876(81)90014-8]
11. Tallal P, Stark RE, Mellits D. The relationship between auditory temporal analysis and receptive language development: evidence from studies of developmental language disorder. Neuropsychologia. 1985;23(4):527-34. [DOI:10.1016/0028-3932(85)90006-5]
12. Sinex DG, Narayan SS. Auditory-nerve fiber representation of temporal cues to voicing in word-medial stop consonants. J Acoust Soc Am. 1994;95(2):897-903. [DOI:10.1121/1.408400]
13. McFayden TC, Baskin P, Stephens JDW, He S. Cortical Auditory Event-Related Potentials and Categorical Perception of Voice Onset Time in Children With an Auditory Neuropathy Spectrum Disorder. Front Hum Neurosci. 2020;14:184. [DOI:10.3389/fnhum.2020.00184]
14. Kraus N, Nicol T. Aggregate neural responses to speech sounds in the central auditory system. Speech Commun. 2003;41(1):35-47. [DOI:10.1016/S0167-6393(02)00091-2]
15. Näätänen R, Picton T. The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. Psychophysiology. 1987;24(4):375-425. [DOI:10.1111/j.1469-8986.1987.tb00311.x]
16. Michalewski HJ, Starr A, Zeng FG, Dimitrijevic A. N100 cortical potentials accompanying disrupted auditory nerve activity in auditory neuropathy (AN): effects of signal intensity and continuous noise. Clin Neurophysiol. 2009;120(7):1352-63. [DOI:10.1016/j.clinph.2009.05.013]
17. Tremblay K, Kraus N, McGee T, Ponton C, Otis B. Central auditory plasticity: changes in the N1-P2 complex after speech-sound training. Ear Hear. 2001;22(2):79-90. [DOI:10.1097/00003446-200104000-00001]
18. Ceponiene R, Alku P, Westerfield M, Torki M, Townsend J. ERPs differentiate syllable and nonphonetic sound processing in children and adults. Psychophysiology. 2005;42(4):391-406. [DOI:10.1111/j.1469-8986.2005.00305.x]
19. Oron A, Szelag E, Nowak K, Dacewicz A, Szymaszek A. Age-related differences in Voice-Onset-Time in Polish language users: An ERP study. Acta Psychol (Amst). 2019;193:18-29. [DOI:10.1016/j.actpsy.2018.12.002]
20. Morris DJ, Tøndering J, Lindgren M. Electrophysiological and behavioral measures of some speech contrasts in varied attention and noise. Hear Res. 2019;373:1-9. [DOI:10.1016/j.heares.2018.12.001]
21. Giraud K, Trébuchon-DaFonseca A, Démonet JF, Habib M, Liégeois-Chauvel C. Asymmetry of voice onset time-processing in adult developmental dyslexics. Clin Neurophysiol. 2008;119(7):1652-63. [DOI:10.1016/j.clinph.2008.02.017]
22. Tremblay KL, Piskosz M, Souza P. Effects of age and age-related hearing loss on the neural representation of speech cues. Clin Neurophysiol. 2003;114(7):1332-43. [DOI:10.1016/s1388-2457(03)00114-7]
23. Dimitrijevic A, Pratt H, Starr A. Auditory cortical activity in normal hearing subjects to consonant vowels presented in quiet and in noise. Clin Neurophysiol. 2013;124(6):1204-15. [DOI:10.1016/j.clinph.2012.11.014]
24. Tremblay KL, Friesen L, Martin BA, Wright R. Test-retest reliability of cortical evoked potentials using naturally produced speech sounds. Ear Hear. 2003;24(3):225-32. [DOI:10.1097/01.AUD.0000069229.84883.03]
25. Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. J Neural Eng. 2018;15(5):056013. [DOI:10.1088/1741-2552/aace8c]
26. Boersma P. Praat, a system for doing phonetics by computer. Glot Int. 2001;5(9):341-5.
27. Dong Y, Gai Y. Speech Perception with Noise Vocoding and Background Noise: An EEG and Behavioral Study. J Assoc Res Otolaryngol. 2021;22(3):349-63. [DOI:10.1007/s10162-021-00787-2]
28. Fritz JB, Elhilali M, David SV, Shamma SA. Auditory attention--focusing the searchlight on sound. Curr Opin Neurobiol. 2007;17(4):437-55. [DOI:10.1016/j.conb.2007.07.011]
29. Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, et al. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron. 2013;77(5):980-91. [DOI:10.1016/j.neuron.2012.12.037]
30. Carter JA, Bidelman GM. Perceptual warping exposes categorical representations for speech in human brainstem responses. Neuroimage. 2023;269:119899. [DOI:10.1016/j.neuroimage.2023.119899]

**Table 1. Model architecture with options**

| Name of the layers | Options |
|---|---|
| Conv_1 | Filters number:16, kernel size: [1,64], strides: [1,1] |
| Conv_2 | Filters number:16, kernel size: [32,1], strides: [1,1] |
| Batch Normalization | |
| Activation | Elu |
| Maxpool_1 | Size: [1,2] |
| Dropout | P = 0.25 |
| Conv_3 | Filter numbers:32, kernel size: [1,32], Stride: [1,1] |
| Batch Normalization | |
| Activation | Elu |
| Maxpool_2 | Size: [1,8] |
| Dropout | P = 0.25 |
| Flatten | |
| Dense | Output size: 3, Activation: "Softmax" |

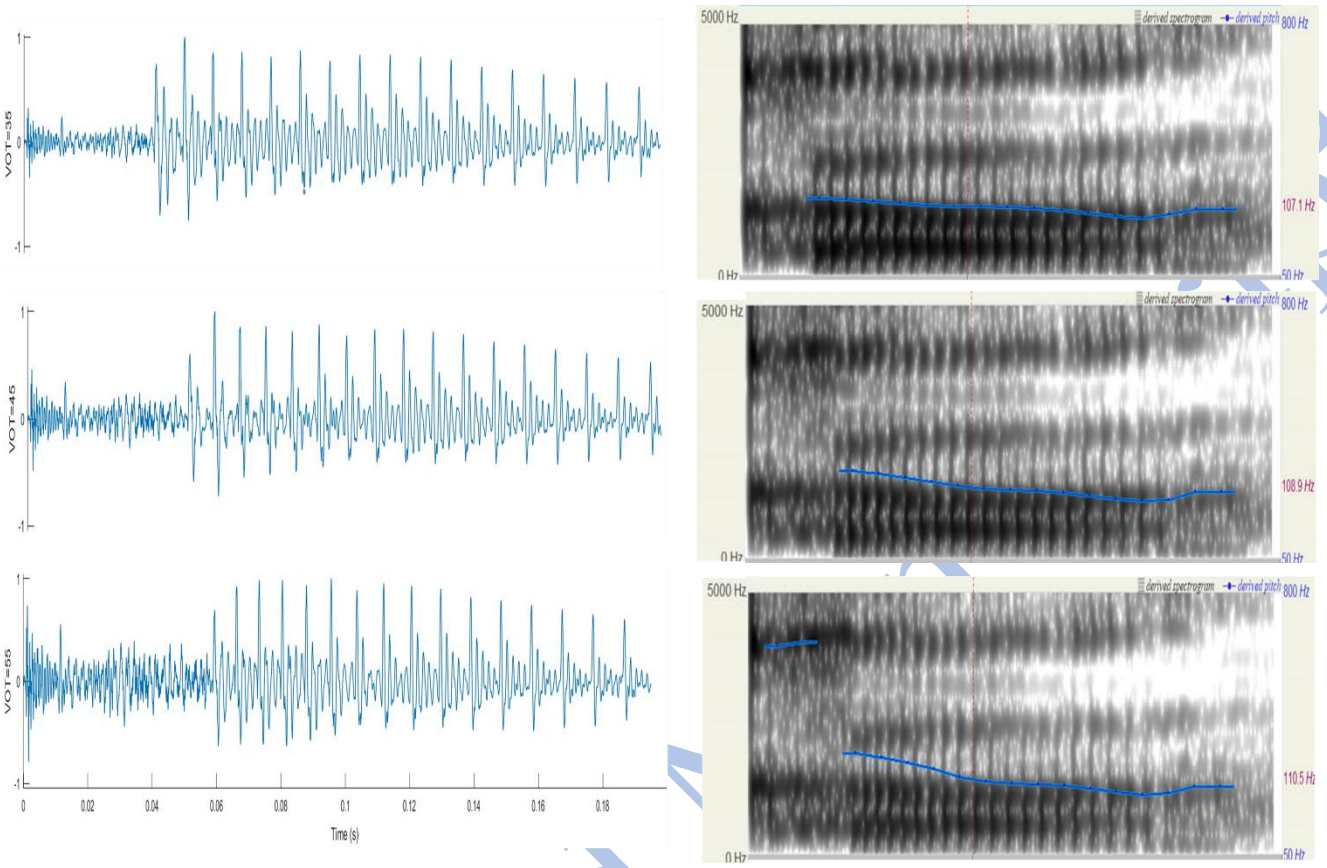


**Fig 1** Temporal waveforms (left) and spectrograms (right) of the three consonant-vowel syllables

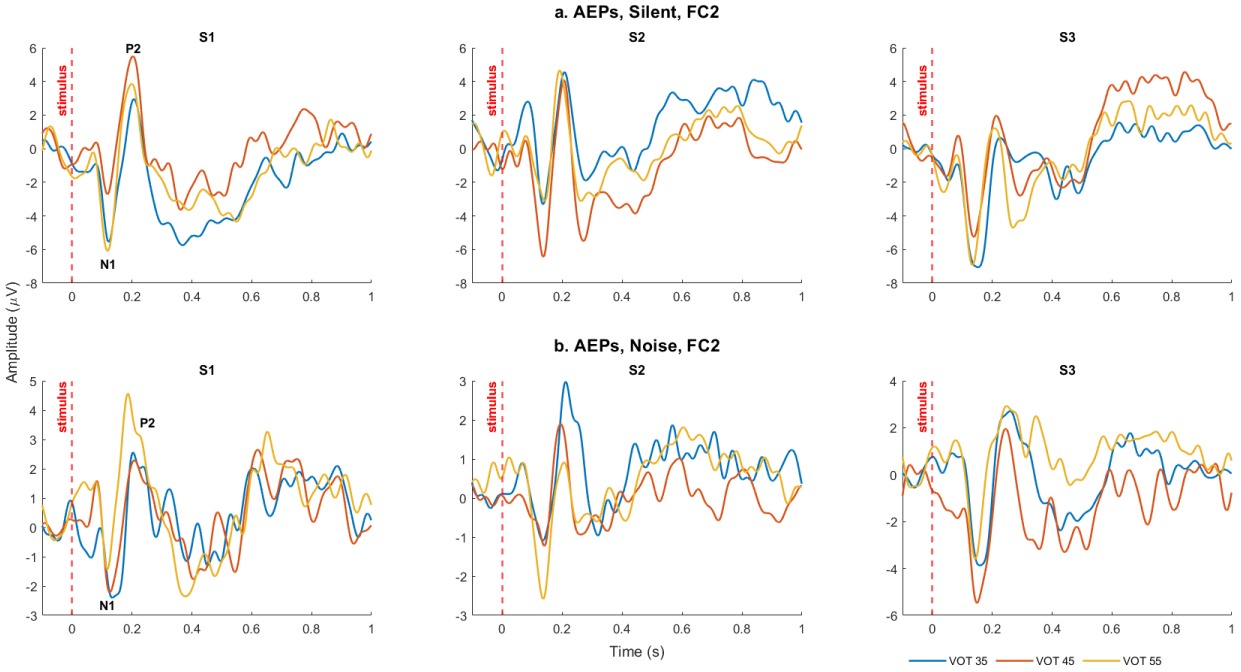


**Fig. 2** Examples of CAEPs obtained with electrode FC2 using silent(a) and noise (b) conditions for three subjects. Note that speech began at time 0, whereas the onset of the epoch was 0.1 s prior to the speech onset, which is the pre-stimulus time. Each CAEP was an average over 160 trials.

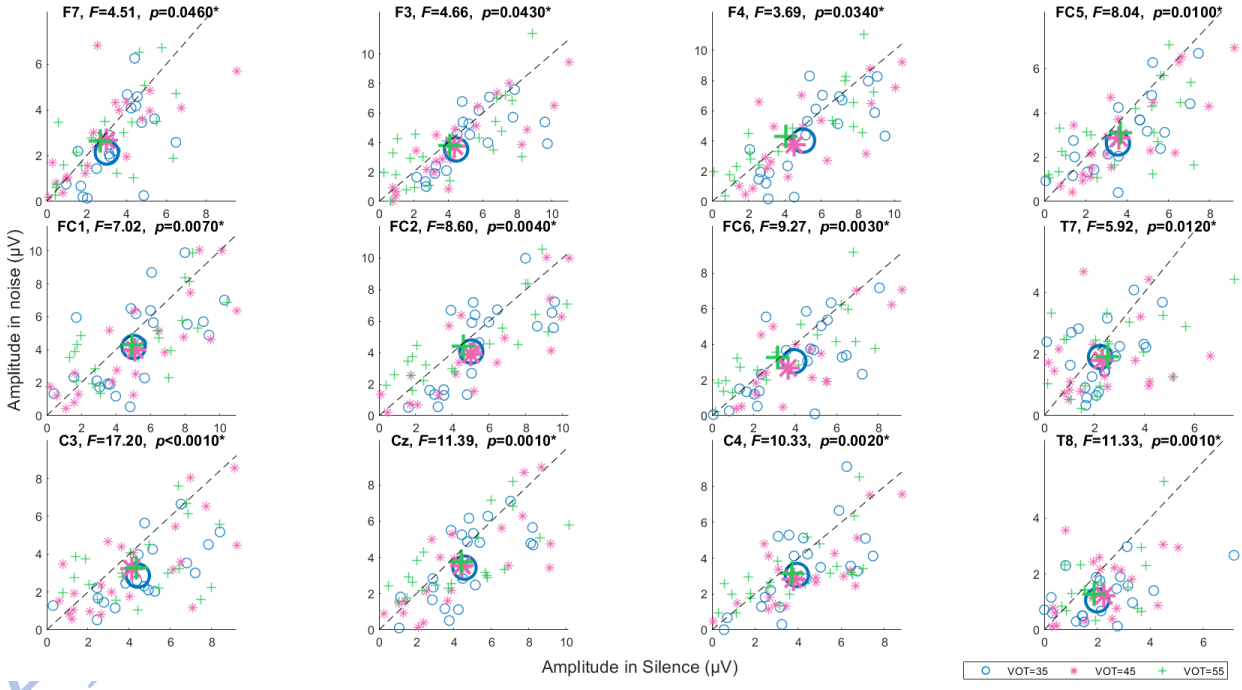


**Fig. 3** Amplitude analysis on N1-P2 complex for all electrodes that showed consistent peaks across subjects. In each scatterplot x-axis shows the absolute complex amplitude for the silent condition and the y-axis shows the absolute complex amplitude to syllables presented in background noise. The channel names and *F* statistics of repeated measure ANOVA with corresponding *p* values are shown above each scatter plot.

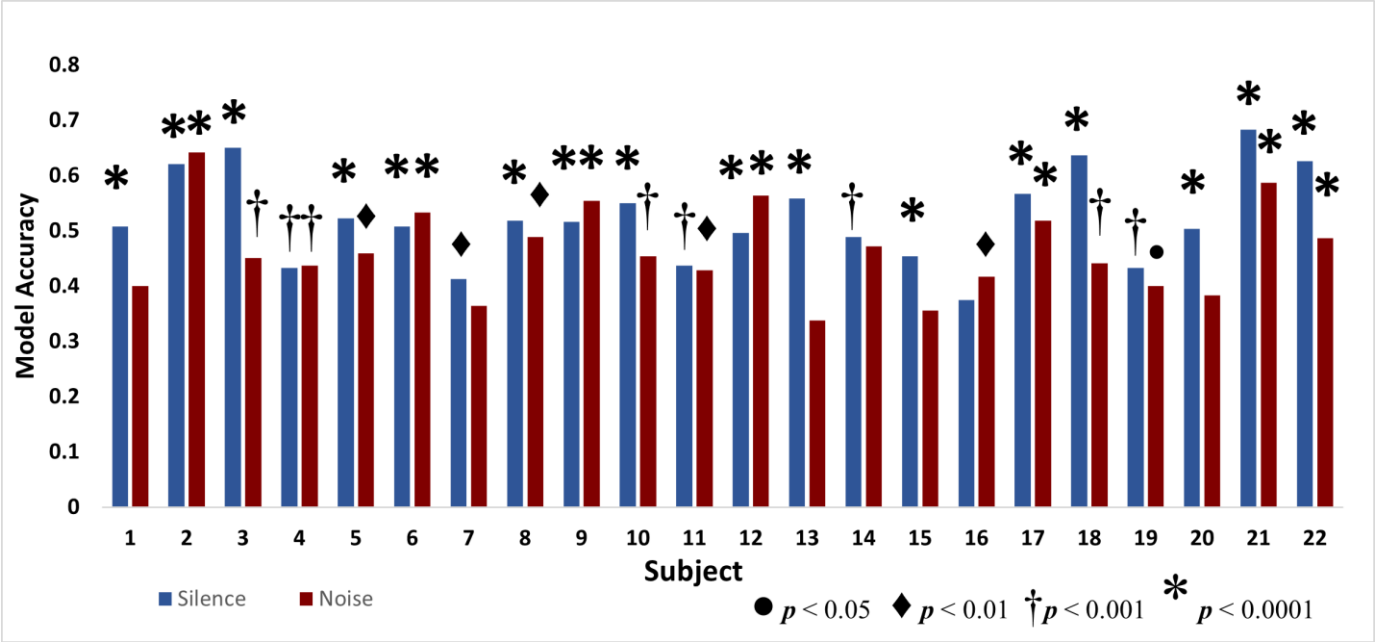


**Fig.4** Classification rates across subjects for Silence and Noise conditions with an asterisk indicating the FDR corrected significance level of the one-sided binomial test of classification accuracy compared to the chance level (33.3%)

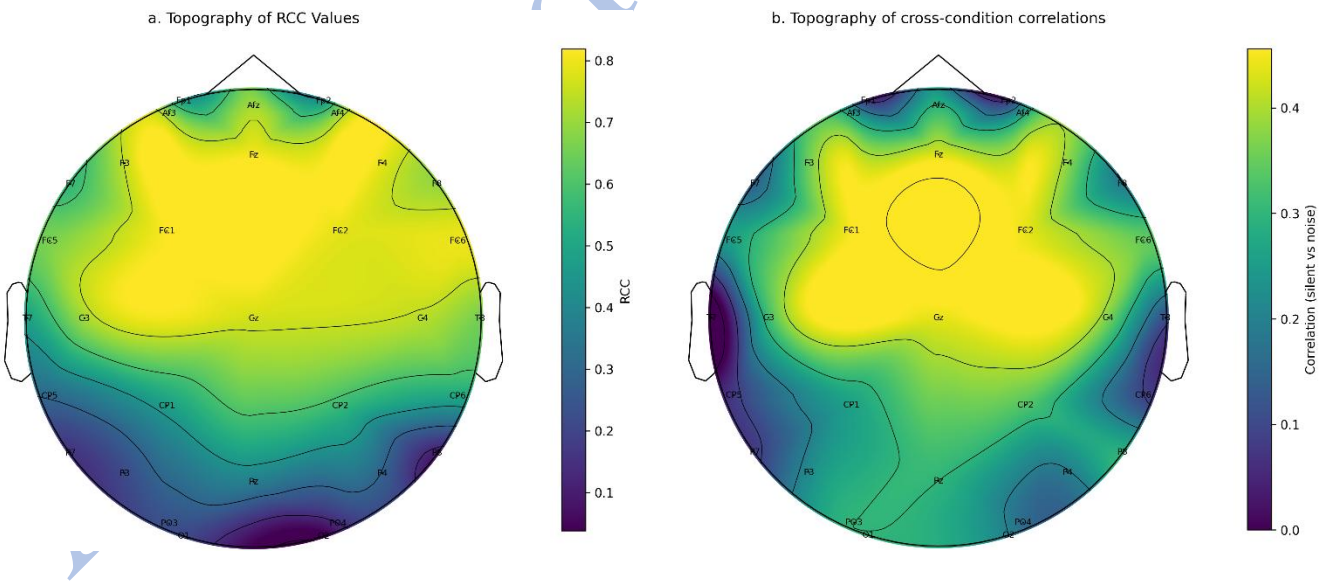


**Fig.5 a** Topography of CAEPs $r_{cc}$ between speech in silence condition and speech in background noise. **b** the topography of the correlation between subject behavioral responses and electrode $r_{cc}$ values across all subjects.

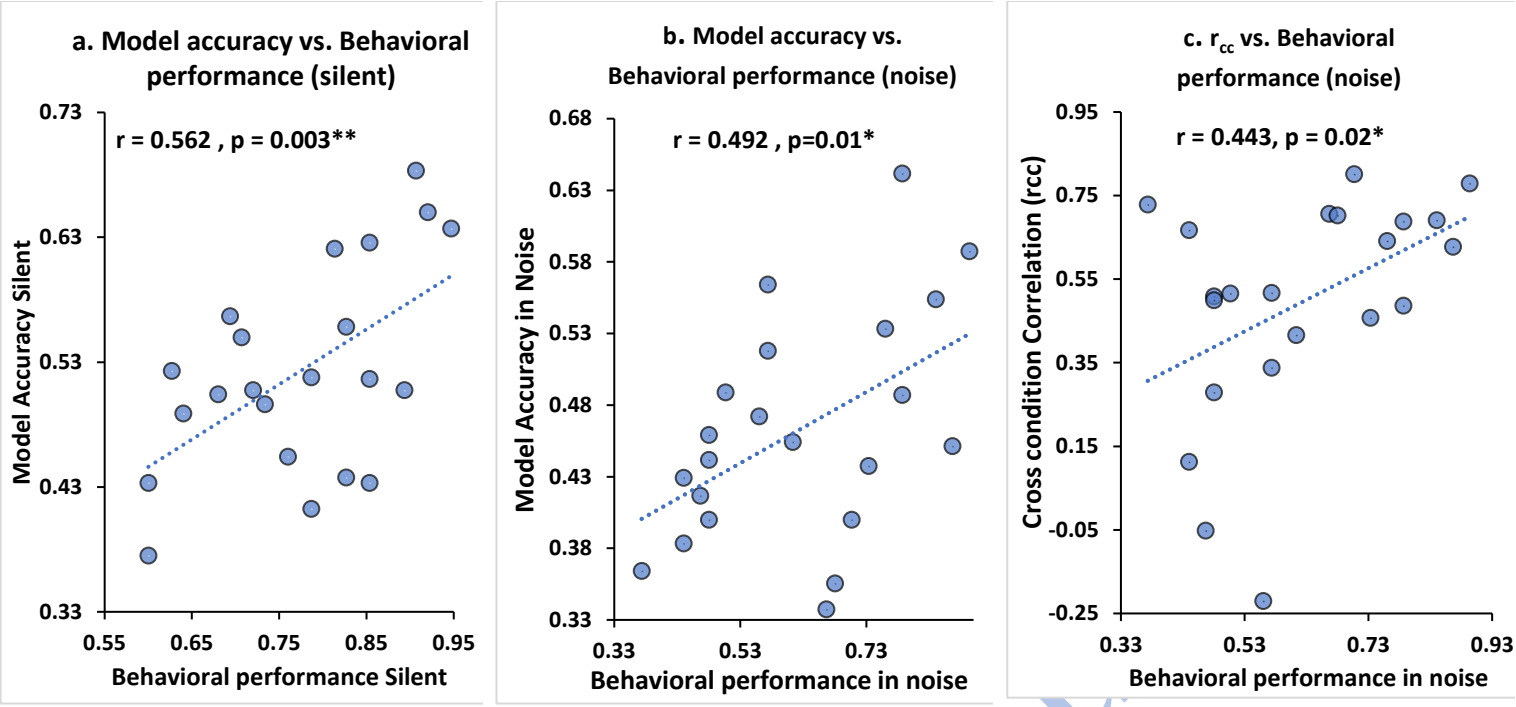


**Fig. 6 a, b** Scatterplots with the linear fit in terms of the correlation between model accuracy and behavioral performance in silence and noise respectively. **c** Scatterplot of $r_{cc}$ values achieved from FC2 electrode and behavioral performance in noise. The statistical values in the figure indicate the quality of the linear fit.