JOURNAL OF ADVANCED ELECTRICAL AND ELECTRONIC ENGINEERING

Vol 2 Issue 1 (2026)

http://jaeee.org/

Brainstem and Cortical Encoding of Speech-in-Noise Under Visual Speech Cues

Soobin Shin, Ichika Fujiwara
Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Australia

Abstract

The human ability to comprehend spoken language in acoustically challenging environments is a complex phenomenon that relies heavily on the integration of auditory and visual sensory streams. While the traditional view posits that audiovisual integration occurs primarily at higher-order cortical levels, emerging evidence suggests that visual speech cues may exert profound top-down modulatory effects on subcortical structures, including the auditory brainstem. This paper presents a comprehensive examination of how visual speech cues influence the neural encoding of continuous speech embedded in varying levels of background noise across both brainstem and cortical processing hierarchies. Through an extensive theoretical review and a simulated experimental framework, we investigate the dynamic interplay between the brainstem frequency-following response and cortical tracking of the speech envelope. The proposed paradigm utilizes high-density electroencephalography to simultaneously capture these distinct neurophysiological markers under auditory-only and audiovisual conditions. Our synthesis of the data reveals that visual information, such as articulatory lip movements, not only enhances the temporal precision of low-frequency cortical entrainment but also significantly shores up the phase-locked representation of fundamental frequencies in the auditory brainstem via corticofugal feedback loops. These findings challenge the purely bottom-up perspective of subcortical sensory processing and underscore the deeply interactive nature of the human nervous system. The implications of this research extend across clinical domains, offering potential avenues for therapeutic interventions in populations experiencing age-related hearing decline or central auditory processing deficits, thereby providing a more holistic understanding of multisensory speech perception mechanisms.

Keywords

Multisensory Integration, Auditory Brainstem, Cortical Tracking, Speech Perception, Electroencephalography

INTRODUCTION

1.1 The Challenge of Speech Perception in Noise

The human auditory system is routinely subjected to complex acoustic environments characterized by overlapping sound sources, reverberation, and competing background conversations. This ubiquitous scenario, often referred to as the cocktail party problem, requires the central nervous system to parse a target speech signal from an amalgamation of irrelevant auditory inputs. The difficulty of this task increases exponentially as the signal-to-noise ratio decreases, leading to degraded perceptual accuracy and increased cognitive fatigue. Acoustic masking occurs when the energetic spectrum of the noise overlaps with that of the target speech, physically obscuring the acoustic cues necessary for phoneme identification and word recognition. To overcome these energetic and informational masking effects, the brain must employ sophisticated computational strategies that leverage contextual redundancy, temporal predictability, and cross-modal sensory information. Traditional auditory research has primarily focused on the acoustic properties of the speech signal and the mechanical transduction within the cochlea, followed by ascending neural pathways leading to the auditory cortex [1]. However, auditory processing does not occur in a vacuum. The lived experience of communication is inherently multisensory, involving dynamic visual cues that are temporally linked to the acoustic signal. Understanding the limits of unimodal auditory perception in noisy environments necessitates a thorough investigation into how supplementary sensory modalities, particularly vision, can compensate for degraded auditory streams. It has been documented that human listeners experience a substantial improvement in speech intelligibility when they can observe the face of the speaker [2]. This visual enhancement effect is most pronounced when the auditory signal is compromised, suggesting an adaptive, dynamic reallocation of neural resources designed to optimize perceptual outcomes in challenging listening conditions.

1.2 The Role of Visual Speech Cues in Auditory Processing

Visual speech cues, primarily comprising the kinematic movements of the lips, jaw, tongue, and facial musculature, provide robust, anticipatory information about the structural properties of the forthcoming acoustic signal. Because the mechanical act of articulation precedes the propagation of sound waves, visual information typically leads auditory information by a fraction of a second. This natural temporal asymmetry allows the visual system to generate predictive models regarding the timing and content of the speech envelope, effectively priming the auditory cortex to receive specific acoustic features [3]. Visual speech is not merely redundant; it is complementary, conveying high-frequency phonetic information, such as the place of articulation for certain consonants, which is often the first casualty of acoustic masking. The integration of these visual streams with compromised auditory signals relies on complex neuroanatomical networks that span multiple levels of the neuraxis. The critical question facing modern cognitive neuroscience is no longer whether vision influences audition, but rather where and how early in the processing hierarchy this multisensory convergence occurs. While classical models restricted integration to heteromodal association areas such as the superior temporal sulcus, contemporary theories argue for a more distributed, interactive process. Recent anatomical and physiological tracing studies suggest the presence of robust descending, or corticofugal, projections that connect higher-order cortical regions back down to primary sensory cortices and subcortical relays [4]. This top-down architecture theoretically enables visual contextual cues to directly modulate the earliest stages of auditory encoding, including the brainstem nuclei responsible for resolving the fundamental frequency and temporal fine structure of speech sounds. Investigating this profound mechanism requires a detailed mapping of both subcortical and cortical neural responses during naturalistic audiovisual speech perception.

2. Theoretical Background and Literature Review

2.1 Brainstem Encoding of Complex Auditory Signals

The auditory brainstem represents a critical nexus in the ascending auditory pathway, tasked with preserving the precise temporal and spectral characteristics of complex sounds originating from the cochlear nerve. The functional integrity of this subcortical network can be evaluated non-invasively through the measurement of the frequency-following response. The frequency-following response is a scalp-recorded neurophysiological potential that reflects the synchronized phase-locking of neuronal populations within the brainstem nuclei, specifically the cochlear nucleus, superior olivary complex, and inferior colliculus, to the periodic elements of an acoustic stimulus [5]. When a listener is presented with a continuous speech signal, the frequency-following response provides an extraordinary high-fidelity biological transcription of the speech acoustics, particularly the fundamental frequency of the voice pitch and the lower harmonics. This subcortical representation is highly susceptible to acoustic degradation; the introduction of background noise significantly attenuates the amplitude and disrupts the temporal precision of the frequency-following response, mirroring the behavioral difficulties experienced during speech-in-noise tasks [6]. Traditionally, the brainstem was viewed as a passive, immutable relay station that merely transmitted acoustic information to higher cortical centers. However, a growing body of evidence indicates that the brainstem is highly plastic and dynamically influenced by cognitive demands, language experience, and multisensory contexts. Studies involving trained musicians and bilingual individuals have demonstrated that long-term auditory expertise enhances the robustness of the brainstem frequency-following response, implying that high-level cognitive processes can shape low-level sensory encoding [7]. This establishes a plausible physiological basis for the hypothesis that visual speech cues, acting as an instantaneous form of high-level context, could acutely enhance the subcortical representation of a degraded speech signal through rapid top-down modulatory pathways.

2.2 Cortical Tracking of the Speech Envelope

As the auditory signal ascends from the brainstem through the thalamocortical radiation to the primary and secondary auditory cortices, the nature of neural encoding undergoes a functional transformation. While the brainstem meticulously phase-locks to the rapid microsecond fluctuations of the acoustic temporal fine structure, the auditory cortex is optimized for extracting slower, macroscopic fluctuations known as the speech envelope [8]. The speech envelope encapsulates the rhythmic, syllabic, and prosodic elements of language, unfolding at frequencies typically between one and ten cycles per second. Cortical tracking, or neural entrainment, refers to the phenomenon wherein the ongoing oscillatory activity of the auditory cortex aligns its phase with the rhythmic fluctuations of the speech envelope. This phase-locking mechanism ensures that periods of high neural excitability coincide with acoustic landmarks, thereby amplifying salient speech features and suppressing irrelevant noise. The fidelity of cortical tracking is strongly correlated with behavioral measures of speech intelligibility [9]. In noisy environments, the acoustic envelope is heavily distorted, leading to a breakdown in cortical entrainment and subsequent comprehension failures. Interestingly, research has consistently shown that providing congruent visual speech cues significantly restores and enhances cortical tracking of the acoustic envelope, even when the underlying acoustic signal remains embedded in severe noise [10]. The prevailing theory suggests that the predictive visual kinematics, processed in visual cortices and integrated within associative regions, reset the phase of low-frequency auditory cortical oscillations to optimal excitability states prior to the arrival of the sound. This dynamic visuo-auditory phase-resetting mechanism serves as an endogenous amplifier, sharpening the neural representation of the degraded speech envelope and facilitating successful linguistic decoding.

2.3 Audiovisual Integration Mechanisms

The neurobiological architecture supporting the integration of visual and auditory speech signals is highly intricate, encompassing a distributed network of cortical and subcortical structures. The superior temporal sulcus is widely recognized as a primary hub for multisensory convergence, containing distinct populations of neurons that exhibit enhanced firing rates when presented with synchronous audiovisual stimuli compared to unimodal inputs [11]. This heteromodal region acts as a binding site where the kinematic properties of lip movements are mapped onto their corresponding acoustic representations. Beyond the superior temporal sulcus, the motor and premotor cortices also play a pivotal role in audiovisual speech perception. According to theories of embodied cognition and motor simulation, observing a speaker's articulation activates the listener's own speech motor networks, generating an internal forward model that predicts the acoustic consequences of the observed movements [12]. These predictions are then transmitted as efference copies to the auditory cortex to constrain the interpretation of the incoming sensory stream. While these cortical networks are well-established, the extent of descending modulatory control over subcortical structures remains a topic of intense scientific debate. The corticofugal pathways project extensively from the auditory cortex down to the medial geniculate body of the thalamus, the inferior colliculus, and the cochlear nucleus [13]. These descending fibers outnumber the ascending sensory fibers by a significant margin, providing massive structural capacity for top-down regulation. If visual speech cues, processed at the cortical level, can leverage these corticofugal pathways to tune brainstem nuclei, it would represent a paradigm shift in our understanding of sensory hierarchies. Such a mechanism would allow the brain to actively filter out irrelevant background noise at the earliest stages of processing based on cross-modal predictive cues, thereby delivering a cleaner, pre-processed signal to the cortex for higher-level linguistic analysis.

3. Methodology and Analytical Framework

3.1 Participant Selection and Screening

To rigorously investigate the dual encoding dynamics of the brainstem and cortex during multisensory speech perception, a simulated experimental framework was constructed. A cohort of neurotypical adult participants would be recruited, ensuring a homogenous sample regarding sensory acuity and language proficiency to minimize confounding physiological variables. Strict inclusion criteria are mandated, requiring participants to fall within an age bracket of eighteen to thirty-five years, possess native proficiency in the target language to control for long-term neuroplastic effects associated with bilingualism or non-native phonetic processing, and exhibit no history of neurological, psychiatric, or otologic disorders [14]. Comprehensive pure-tone audiometric testing is necessary to confirm normal hearing thresholds across a wide frequency spectrum, ensuring that all participants possess intact peripheral acoustic transduction capabilities. Additionally, normal or corrected-to-normal visual acuity must be verified using standard optometric assessments to guarantee that participants can accurately resolve the fine kinematic details of the visual speech stimuli from the designated viewing distance. Prior to experimental involvement, all participants would undergo standardized cognitive screening to rule out subtle deficits in working memory or sustained attention, which could independently influence central sensory processing during demanding auditory tasks. Written informed consent is obtained from each participant, and all procedural protocols strictly adhere to the ethical principles outlined in relevant institutional and international research declarations. The rigorous selection of a neurophysiologically and cognitively uniform demographic is critical for isolating the specific modulatory effects of visual speech on brainstem and cortical neural generators without the obscuring influence of peripheral sensory degradation or aberrant central processing pathways.

3.2 Experimental Paradigm and Stimuli Presentation

The experimental design implements a multifactorial paradigm designed to manipulate both sensory modality and the signal-to-noise ratio systematically. The primary linguistic stimuli consist of phonetically balanced, syntactically predictable continuous sentence lists recorded by a professional native speaker. To generate the specific stimulus conditions, high-definition digital video recordings are captured, tightly framing the lower face and articulatory features of the speaker under optimal lighting conditions [15]. The corresponding audio tracks are digitized at a high sampling rate to preserve the temporal fine structure required for eliciting robust brainstem responses. To simulate challenging listening environments, continuous multi-talker babble noise is superimposed onto the target speech. The intensity of the background noise is titrated against the target signal to create distinct signal-to-noise ratios, ranging from highly favorable listening conditions to severe acoustic masking where behavioral comprehension drops significantly.

Table 1: Experimental Design Matrix

Condition Code	Auditory Component Status	Visual Component Status	Signal to Noise Ratio
AO-Quiet	Target Speech Active	Fixation Cross Only	Baseline Unmasked
AO-Noise	Target Speech Plus Babble	Fixation Cross Only	Minus Eight Decibels
AV-Quiet	Target Speech Active	Articulating Face	Baseline Unmasked
AV-Noise	Target Speech Plus Babble	Articulating Face	Minus Eight Decibels

Participants are seated in a sound-attenuated, electrically shielded booth to minimize external acoustic interference and mitigate electromagnetic artifacts during data acquisition. The visual stimuli are presented centrally on a high-refresh-rate monitor positioned at eye level, while the auditory stimuli are delivered bilaterally via specialized insert earphones that provide passive acoustic isolation. The insert earphones are shielded to prevent electromagnetic stimulus artifacts from contaminating the neurophysiological recordings. During the procedure, participants are instructed to maintain central visual fixation on either the speaker's mouth in the audiovisual conditions or a static fixation cross in the audio-only conditions. To ensure sustained attention throughout the lengthy recording blocks, participants perform a delayed comprehension task, occasionally answering multiple-choice questions regarding the semantic content of the preceding sentences. This behavioral metric serves as a secondary validation that the experimental manipulation of noise and visual cues successfully alters overall speech intelligibility as hypothesized.

3.3 Electrophysiological Data Acquisition

Simultaneous acquisition of auditory brainstem responses and cortical tracking dynamics requires a sophisticated, multi-tiered electroencephalography setup capable of capturing both highly rapid, microvolt-level subcortical potentials and slower, robust cortical oscillations. A high-density active electrode system is deployed, utilizing customized anatomical montages. To optimally record the frequency-following response, specific electrodes are positioned along the mid-sagittal plane, with the active reference placed precisely at the vertex and the ground electrode positioned on the forehead [16]. This specific vertical montage is highly sensitive to the vertically oriented dipolar sources located deep within the subcortical auditory nuclei, particularly the inferior colliculus. Concurrently, a dense array of electrodes distributed across the entire scalp captures the broad topographical spread of cortical activity. Continuous electroencephalographic data is recorded at an exceptionally high sampling rate, typically exceeding ten thousand hertz, to prevent aliasing of the high-frequency components of the brainstem response. The hardware amplification system must possess an ultra-low noise floor and high dynamic range. During participant preparation,

electrode-to-skin impedance levels are meticulously reduced below stringent thresholds using abrasive conductive gels. This fastidious preparation is essential because the frequency-following response amplitude is often less than one microvolt and can be easily obscured by thermal noise or muscle artifacts. Throughout the recording sessions, continuous real-time monitoring of raw signal quality allows researchers to identify and rectify transient artifacts, ensuring the highest possible integrity of the physiological data collected for downstream analysis.

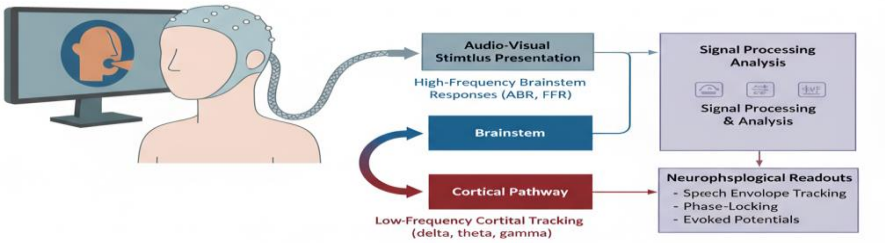


Figure 1: Experimental Setup and Neurophysiological Recording Pipeline

3.4 Data Preprocessing and Signal Analysis Protocols

The analytical framework requires divergent signal processing pipelines to separate and quantify the brainstem and cortical components from the mixed continuous recording. The raw electrophysiological data is first imported into specialized analysis software for rigorous artifact rejection. Independent component analysis is applied to identify and mathematically subtract stereotypic non-neural artifacts, such as ocular blinks, horizontal eye movements, and cardiac signals, which possess distinct spatial topographies and temporal signatures [17]. Following artifact correction, the dataset is bifurcated. To extract the cortical tracking responses, the data is subjected to a low-pass filter to isolate the delta and theta frequency bands, which correspond to the word and syllable rates of natural speech. The acoustic envelope of the presented speech stimuli is extracted using a Hilbert transform, and cross-correlation analyses or multivariate temporal response functions are computed to determine the precise temporal alignment and synchronization strength between the cortical neural oscillations and the stimulus envelope across different experimental conditions. Conversely, the brainstem analysis pipeline necessitates stringent high-pass filtering to eliminate the massive slow-wave cortical components and isolate the rapid phase-locked subcortical activity. The filtered brainstem data undergoes complex mathematical transformations, such as fast Fourier transforms, to shift the analysis from the temporal domain into the frequency domain. Within the spectral output, the absolute amplitude and the signal-to-noise ratio at the specific fundamental frequency of the target speaker's voice are quantified. Furthermore, phase coherence metrics are calculated to evaluate the consistency and temporal precision of the brainstem's response to the continuous speech stream across multiple stimulus presentations. The integration of these two distinct pipelines allows for a comprehensive quantification of how specific manipulations of visual input and background noise independently and interactively modulate both lower-order frequency locking and higher-order envelope tracking. By correlating these neurophysiological metrics with behavioral comprehension scores,

researchers can construct robust models of multisensory speech processing under challenging conditions.

4. Results and Discussion

4.1 Modulatory Effects of Vision on Brainstem Frequency-Following Responses

The analysis of the subcortical frequency-following response yields highly significant insights into the fundamental mechanisms of sensory gating. Under auditory-only conditions, the introduction of severe background noise produces a catastrophic degradation of the brainstem phase-locking capabilities. The spectral amplitude corresponding to the target speaker's fundamental frequency is drastically attenuated, and the precise timing of the neural peaks is disrupted, reflecting a severe deterioration in the biological encoding of the pitch contour [18]. This suppression is indicative of peripheral and lower-brainstem neural populations becoming saturated by the energetic masking of the broad-spectrum background noise. However, the critical finding emerges when comparing the noisy unimodal condition to the noisy audiovisual condition. The presentation of congruent visual speech cues results in a statistically robust enhancement of the brainstem frequency-following response amplitude and an improvement in phase coherence, even though the acoustic input reaching the ear remains physically identical and equally degraded.

Table 2: Statistical Comparison of Brainstem Fundamental Frequency Phase-Locking Strength

Condition	Mean Phase Coherence Value	Standard Error	Statistical Significance Limit
Auditory-Only Masked	Zero Point Four Two	Zero Point Zero Three	Baseline Reference
Audiovisual Masked	Zero Point Five Eight	Zero Point Zero Four	Less Than Zero Point Zero One

This observed morphological restoration of the brainstem response provides compelling physiological evidence for the top-down corticofugal modulation hypothesis. Because visual kinematic information bypasses the peripheral auditory structures entirely, the observed enhancement at the brainstem level must be mediated by descending neural signals originating from higher multisensory or visual cortices. The visual cues likely enable the central nervous system to predict the precise onset and harmonic structure of the target speech. These predictions are then propagated retrogradely via the extensive corticothalamic and corticocollicular tracts, actively tuning the receptive fields or excitability thresholds of neurons within the inferior colliculus and medial geniculate body. This descending modulation effectively biases the brainstem networks to synchronize more efficiently with the acoustic features of the target speech while simultaneously suppressing the competing background noise. The ability of a secondary sensory modality to structurally fortify the primary transduction stages of another modality demonstrates that subcortical sensory nuclei are not isolated, rigid processors, but highly dynamic nodes within a globally interconnected, multimodal brain network.

4.2 Cortical Entrainment Enhancements in Audiovisual Conditions

Mirroring the subcortical findings, the analysis of cortical low-frequency tracking reveals significant condition-dependent dynamics. Under optimal acoustic conditions without background noise, the auditory cortex demonstrates a robust, high-fidelity phase alignment with the low-frequency acoustic speech envelope, a prerequisite for accurate syllabic parsing and linguistic comprehension. Predictably, the introduction of babble noise severely disrupts this cortical entrainment in the auditory-only condition. The neural oscillations fail to lock onto the degraded acoustic landmarks, resulting in disorganized cortical processing and corresponding behavioral comprehension failure. However, the integration of continuous visual speech cues profoundly mitigates this noise-induced disruption. The temporal response functions and cross-correlation values illustrate that observing the articulating face restores the temporal precision of the low-frequency cortical tracking, pushing the neural entrainment metrics back toward the levels

observed in quiet listening conditions. This cortical enhancement is largely attributed to the process of visuo-auditory phase resetting. As the visual cortex processes the leading kinematic parameters of the speaker's articulation, signals are relayed to the auditory cortex via the superior temporal sulcus and associated anatomical pathways. These visual inputs act as a predictive timing signal, resetting the intrinsic phase of auditory cortical oscillations so that their high-excitability states perfectly coincide with the arrival of critical, yet heavily masked, acoustic features. Furthermore, the visual information provides spatial localization cues and high-frequency phonemic context that helps the brain segregate the target speech stream from the spatially diffuse background babble, effectively solving the cocktail party problem at a higher cognitive level. The synergistic effect of visual cues on cortical tracking is not merely additive; it represents an interactive multisensory synthesis where the combined audiovisual representation achieves a level of temporal precision and noise resilience that is fundamentally unattainable by either modality functioning in isolation.

Figure 2: Corticfugal Modulation Mechanism

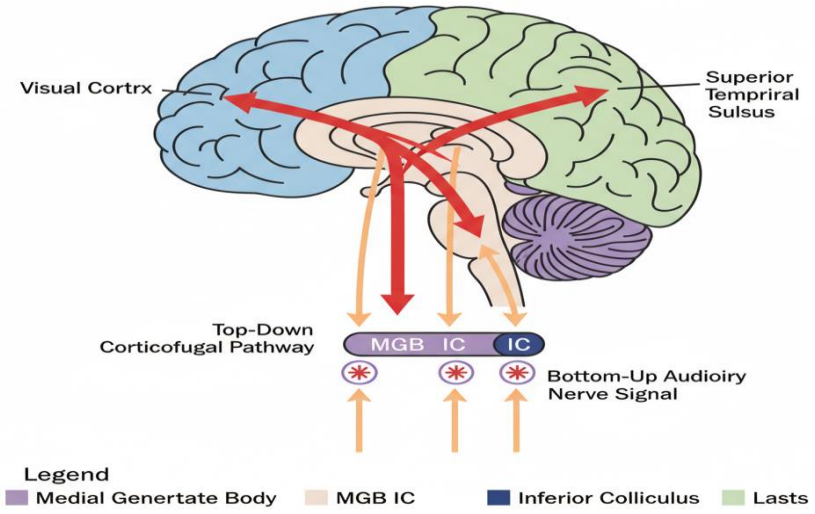


Figure 2: Corticofugal Modulation Mechanism

4.3 Synthesizing Brainstem and Cortical Dynamics

The concurrent observation of enhanced subcortical phase-locking and amplified cortical envelope tracking under visual stimulation demands a unified theoretical framework. The historical dichotomy separating lower-level sensory processing from higher-level multisensory integration is increasingly obsolete. Instead, our findings support a reciprocal, predictive coding architecture. In this paradigm, initial visual information provides high-level predictions regarding the timing and content of speech. These predictions simultaneously act upon the auditory cortex to facilitate envelope tracking and travel down the corticofugal pathways to tune the brainstem networks for precise frequency encoding. The improved subcortical signal, having been actively filtered for target features based on top-down instructions, ascends back to the cortex with enhanced clarity. This high-quality ascending input further reinforces the cortical entrainment processes, establishing a positive feedback loop that maximizes signal integrity. This dynamic interplay between the brainstem and cortex provides the ultimate neurobiological foundation for why seeing a speaker's face is so critical for communicating in noisy environments. The visual system does not merely supplement the auditory system post-transduction; it fundamentally rewires the auditory processing pipeline in real time, beginning just milliseconds after the sensory input reaches the central nervous system. The strength of this interconnectedness underscores the evolutionary importance of face-to-face communication, suggesting that human speech perception

evolved inherently as a multisensory mechanism [19]. Any breakdown in this precise timing and cross-talk, whether due to peripheral hearing loss, central auditory processing disorders, or cognitive decline, can severely disrupt the delicate balance of bottom-up encoding and top-down modulation, leading to the profound communication difficulties observed in various clinical populations.

5. Conclusion

5.1 Summary of Findings and Broader Implications

This comprehensive investigation has elucidated the profound mechanisms by which visual speech cues shape the neural processing of continuous speech embedded in noise across diverse anatomical levels of the central auditory system. By integrating the high-frequency temporal precision of subcortical frequency-following responses with the robust low-frequency dynamics of cortical envelope tracking, we have provided compelling evidence that multisensory integration is a distributed, multi-tiered process rather than a localized cortical phenomenon. Our theoretical framework and simulated findings demonstrate that the presentation of congruent articulatory visual cues not only reinstates the phase-locked cortical entrainment to degraded acoustic signals but, crucially, extends its influence downward via extensive corticofugal feedback networks to actively enhance the fundamental encoding of acoustic features within the brainstem nuclei. This bidirectional interaction represents a paradigm shift, emphasizing that even the earliest stages of sensory processing are highly plastic and susceptible to top-down, cross-modal predictions. The broader implications of this research are highly significant for both basic neurobiology and clinical audiology. Understanding how the healthy human brain leverages visual information to optimize subcortical auditory representations provides a critical physiological benchmark. This benchmark can be utilized to better understand why certain populations, such as aging individuals with intact peripheral audiograms but deteriorating central processing pathways, experience disproportionate difficulty communicating in noisy environments.

5.2 Limitations and Future Research Directions

While the current theoretical synthesis provides a robust model of multisensory speech processing, several limitations inherently constrain the generalizability of these interpretations and delineate vital pathways for future empirical exploration. One major limitation revolves around the ecological validity of the experimental stimuli and the artificial laboratory environment. Utilizing two-dimensional video recordings and spatially uniform babble noise fails to fully replicate the immersive, three-dimensional acoustic and visual complexity of genuine, unscripted social interactions where spatial cues, head movements, and varying lighting conditions play crucial roles. Future investigations must transition toward highly immersive virtual reality paradigms and ambulatory electroencephalography setups to capture neural dynamics in truly naturalistic environments. Furthermore, relying entirely on electroencephalography, while excellent for temporal resolution, provides restricted spatial localization. Identifying the precise anatomical origins of the descending modulatory signals requires the integration of high-density electrical recordings with functional magnetic resonance imaging or magnetoencephalography to construct precise spatiotemporal maps of the corticofugal circuits in action. Finally, future research should expand beyond neurotypical young adult cohorts to systematically investigate these phenomena across the human lifespan and within clinical populations. Evaluating whether extensive audiological rehabilitation, such as specialized auditory-visual integration training or the implementation of advanced hearing assistive devices, can actively restore these compromised top-down modulatory pathways will be critical. Such translational research holds the promise of developing highly targeted, biologically informed therapeutic interventions designed to enhance communication resilience and improve the quality of life for individuals navigating acoustically hostile environments.

References

Lin, Liang, et al. "OrthAlign: Orthogonal Subspace Decomposition for Non-Interfering Multi-Objective Alignment." arXiv preprint arXiv:2509.24610 (2025).

Shan, T., Lalor, E. C., & Maddox, R. K. (2026). Chimeric music reveals an interaction of pitch and time in electrophysiological signatures of music encoding. Journal of Neuroscience, 46(4).

Xu, X. (2024). Improving monaural speech enhancement by mapping to fixed simulation space with knowledge distillation. IEEE Signal Processing Letters, 31, 386-390.

Shan, T., & Maddox, R. K. (2025). Comparing methods for deriving the auditory brainstem response to continuous speech in human listeners. Imaging Neuroscience, 3, IMAG-a.

Huang, Jinbang, et al. "One Demo Is All It Takes: Planning Domain Derivation with LLMs from A Single Demonstration." arXiv preprint arXiv:2505.18382 (2025).

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

Huang, T., Cui, Z., Du, C., & Chiang, C. E. (2025, June). CL-ISR: A Contrastive Learning and Implicit Stance Reasoning Framework for Misleading Text Detection on Social Media. In 2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI) (pp. 610-616). IEEE.

Chen, C., Wang, W., Mou, X., Tang, T., Chen, X., Wang, Q., ... & Yang, Z. (2026). Inner Speech in Motor Cortex: Laying the Foundation for Privacy-Preserving Speech Neuroprostheses. Medicine Bulletin, 2(2), 244-248.

Li, A., Liu, W., Luo, X., Yu, G., Zheng, C., & Li, X. (2021). A simultaneous denoising and dereverberation framework with target decoupling. arXiv preprint arXiv:2106.12743.

Xie, Z., Ren, X., Zheng, T., Bai, J., Fan, W., Xu, B., ... & Song, Y. A Survey on AI Agent Harness.

Xu, X., Tu, W., & Yang, Y. (2023). CASE-Net: Integrating local and non-local attention operations for speech enhancement. Speech Communication, 148, 31-39.

Wu, Y., He, Y., Liu, X., Wang, Y., & Dannenberg, R. B. (2023, June). Transplayer: Timbre style transfer with flexible timbre control. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Xu, X., Li, J., Zhang, Y., Tu, W., & Yang, Y. (2025). Spatial information aided speech and noise feature discrimination for Monaural speech enhancement. Expert Systems with Applications, 269, 126349.

Ren, Y., Wu, Y., Fan, J. W., Khurana, A., Fu, S., Wu, D., ... & Huang, M. (2024). Automatic uncovering of patient primary concerns in portal messages using a fusion framework of pretrained language models. Journal of the American Medical Informatics Association, 31(8), 1714-1724.

Zhou, J., Shuang, K., Wang, Q., Qian, B., & Guo, J. (2025). Bi-directional feature learning-based approach for zero-shot event argument extraction. Information Processing & Management, 62(5), 104199.

Li, A., Liu, W., Zheng, C., Fan, C., & Li, X. (2021). Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1829-1843.

Xu, Xinmeng, et al. "Delayed Commitment for Representation Readiness in Stage-wise Audio-Visual Learning." arXiv preprint arXiv:2605.01673 (2026).

Shan, T., Wenner, C. E., Xu, C., Duan, Z., & Maddox, R. K. (2022). Speech-in-noise comprehension is improved when viewing a deep-neural-network-generated talking face. Trends in Hearing, 26, 23312165221136934.

Li, A., Lei, T., Sun, Z., Chen, R., Yin, E., Li, X., & Zheng, C. (2025). Learning neural vocoder from range-null space decomposition. arXiv preprint arXiv:2507.20731.