Proceedings of ICEIPI 2025 Symposium: Reimagining Society: AI's Role in Cultural Transformation and Learning Environments
DOI: 10.54254/2753-7048/2025.BO25686

Phonemic Differences in the McGurk Effect in Mandarin Chinese: An Exploration Based on Acoustic Features and Audiovisual Integration Mechanisms

Shuchen Yao[1*†], Xinyu Zhou[2†], Zixi Zhao[3]

[1]College of Humanities, Zhejiang Normal University, Jinhua, China
[2]Department of Chinese (Zhuhai), Sun Yat-sen University, Zhuhai, China
[3]Desheng School (International), Guangdong, China
**Corresponding Author. Email: LookOut4ASec@outlook.com*
[†]These authors contributed equally to this work and should be considered co-first authors.

Abstract. This study examines phonemic differences in the McGurk effect in Mandarin Chinese, emphasizing acoustic features and audiovisual integration. Fifty-nine native Mandarin speakers completed three tasks: Basic Phoneme Selection, Tone-Consonant Separation Judgment, and Complex Articulation Contrasts. A one-way ANOVA revealed a significant main effect of task type on accuracy ($F=20.251$, $p<0.001$). Accuracy was highest in Task 1 ($M=0.85$), followed by Task 3 ($M=0.65$), with Task 2 showing the lowest accuracy ($M=0.45$), indicating challenges in resolving tone-consonant conflicts. Fricative/stop combinations (e.g., /f/ vs. /p^h/) elicited a higher fusion perception rate than pure stops ($F=7.144$, $p=0.026$), attributed to acoustic ambiguity and visual complementarity. Labiodental sounds (e.g., /f/) demonstrated significantly higher fusion rates ($M=0.16$, $SD=0.12$) than non-labiodentals ($M=0.03$, $SD=0.04$), highlighting visual salience (e.g., lip-teeth contact) in perceptual integration. Findings suggest Mandarin speakers' heightened sensitivity to segmental conflicts, potentially influenced by tonal language structures. These results inform speech synthesis optimization (e.g., lip-sync enhancement) and cross-linguistic audiovisual algorithm design. Future research should integrate neuroimaging to explore neural mechanisms and dialectal impacts on multimodal processing.

Keywords: McGurk effect, audiovisual integration, Mandarin Chinese, multimodal perception

1. Introduction

In 1976, McGurk and MacDonald proposed the classic audiovisual speech perception effect. the academic community has gradually revealed the core role of multimodal integration in speech comprehension. A large number of studies (such as [1]) have confirmed that visual cues (such as lip movements) significantly affect auditory perception, especially in the presence of audiovisual conflict, subjects tend to rely on visual information to produce illusory perception [2]. Crosslinguistic studies further show that the intensity of the McGurk effect is language-specific: for

DOI: 10.54254/2753-7048/2025.BO25686

example, Finnish and Japanese subjects have significantly different effects under Japanese stimuli [3], and cultural factors (such as language phonological structure) are considered to be key moderating variables [4].

In the Chinese context, the tone system and syllable structure of Mandarin gives it a unique perception mechanism [5]. Liu Shun, Chen Xuemei, and Wang Suiping analyzed the influence of the Chinese tone system and syllable structure on speech comprehension, pointing out that syllable structure affects the way listeners process speech [6]; Zhang Yong, Kuhl PK, Imada T, Kotani M, and Tohkura Y. drew an important conclusion in their study: the behavior of Mandarin speakers to classify speech according to tone and syllable structure enables them to perceive speech more effectively.

Existing Chinese McGurk studies mainly focus on consonants (such as stops/frictions) and simple finals (such as /a/ and /i/), while the audiovisual integration mechanism of complex phoneme combinations (such as nasal codas and rounded vowels) has not been fully explored. Based on this, we will expand the research on phonemes and tones, incorporate complex phoneme combinations (such as affricates), and construct a complete audiovisual integrated speech map of Chinese.

Our research is based on the audiovisual interaction hypothesis: phonemes with complex acoustic features (such as fricatives) or significant visual cues (such as labiodentals) will trigger a stronger McGurk effect.

Specifically: Fricative/stop combinations (such as /f/ vs. /p^h/) may enhance the perceptual integration of audiovisual conflict due to their spectral overlap and visual dynamic complementarity (combination of fricatives and plosives); Labiodentals (such as /f/) may preferentially trigger visual-dominated perceptual integration due to the high visibility of the articulation site (lip-tooth contact).

2. Method

2.1. Participants

A total of 59 native Mandarin speakers (18 males, 41 females) aged 15–21 years participated in the experiment. Participants were recruited based on the following criteria: (1) native proficiency in Mandarin Chinese, (2) no history of hearing impairments or attention deficits, and (3) enrollment in high school or university to ensure standardized educational backgrounds and cognitive consistency. All talkers had passed the National Mandarin Proficiency Test (Level 1B or higher), ensuring articulatory clarity in stimulus recordings. Ethical approval was obtained from the Institutional Review Board (IRB2023001), and written informed consent was secured from all participants.

2.2. Materials and design

2.2.1. Stimuli

Baseline stimuli: Audiovisually congruent videos featuring native Mandarin speakers articulating phonemes. Videos focused on lip movements, with nasal articulations (e.g., /m/, /n/) additionally displaying nasal cavity visibility.

Conflict stimuli: Audiovisually incongruent videos (e.g., auditory /pa/ paired with visual /ka/).

2.2.2. Control variables

Talker gender balance (50% male, 50% female).

Standardized video resolution (1080p) and audio sampling rate (44.1 kHz).

DOI: 10.54254/2753-7048/2025.BO25686

Articulatory features (e.g., lip protrusion, nasal airflow) quantified via frame-by-frame analysis.

2.3. Task design

Three experimental tasks were administered using a forced-choice paradigm:

1. Task 1 (Basic Phoneme Selection): Participants selected perceived phonemes from auditory input. Example pairs included:

[pʰʌ51] (怕 "fear") vs. [pʌ51] (爸 "father") (aspirated vs. unaspirated plosives).

[fʌ55] (发 "issue") vs. [pʰʌ55] (趴 "crawl") (fricative vs. aspirated plosive).

[sʌ214] (洒 "sprinkle") vs. [ʂʌ214] (傻 "foolish") (alveolar vs. retroflex fricatives).

2. Task 2 (Tone-Consonant Separation Judgment): Participants evaluated congruency between auditory consonants and visual lip movements. Example pairs included:

[tʌ55] (搭 "attach") vs. [tʰʌ214] (塔 "tower") (unaspirated vs. aspirated plosives with tonal contrast).

[mʌ55] (妈 "mother") vs. [mʌ51] (骂 "scold") (identical consonant with highlevel vs. falling tone).

3. Task 3 (Complex Articulation Contrasts): Participants compared phonemes with intricate articulatory features. Examples included:

Nasal coda contrasts: [pan55] (班 "class") vs. [pɑŋ55] (帮 "help") (anterior vs. posterior nasal codas).

Rounded vs. unrounded vowels: [li214] (李 "plum") vs. [ly214] (吕 "surname Lü").

Aspiration contrasts in sentences: 他跑[pʰɑo214]得如兔子一样快 ("He runs as fast as a rabbit") vs. 他抱[pɑo51]得如兔子一样快 ("He holds as fast as a rabbit").

2.4. Procedure

The experiment was conducted between February 24 and March 2, 2025, across three locations:

1. Phonetics Laboratory, Department of Chinese Language and Literature, Sun Yatsen University (Zhuhai Campus).

2. Language Experiment Center, School of Foreign Studies, Zhejiang Normal University.

3. Information Technology Classroom, Desheng School (International Division), Shunde, Guangdong.

All sessions were conducted in sound-attenuated, distraction-free environments. Participants completed tasks sequentially, with stimuli presented in randomized order.

3. Analyses

The provided analyses and graphs offer a comprehensive overview of the impact of different task types on speech perception accuracy, particularly focusing on the McGurk effect, which illustrates the integration of auditory and visual information in speech perception.

3.1. The significant impact of task types on accuracy

The oneway ANOVA indicates that the type of task has a highly significant main effect on accuracy ($F=20.251$, $p=0.000$). Specifically, the results are as follows: Task 1 (Basic Phoneme Selection): The average accuracy is the highest ($M=0.85$), indicating that participants perform well overall in distinguishing between the voiced and voiceless contrasts of initial consonants and in identifying flat versus retroflex sounds. Task 2 (Tone-Consonant Separation Judgment): The accuracy significantly

DOI: 10.54254/2753-7048/2025.BO25686

declines to M=0.45, suggesting that participants face considerable challenges when dealing with conflicts between tone and consonant simultaneously. Task 3 (Complex Pronunciation Comparison): The accuracy improves to M=0.65, yet remains lower than that of Task 1, possibly due to the increased perceptual load from complex phoneme combinations (such as nasal finals and rounded vowel sounds).

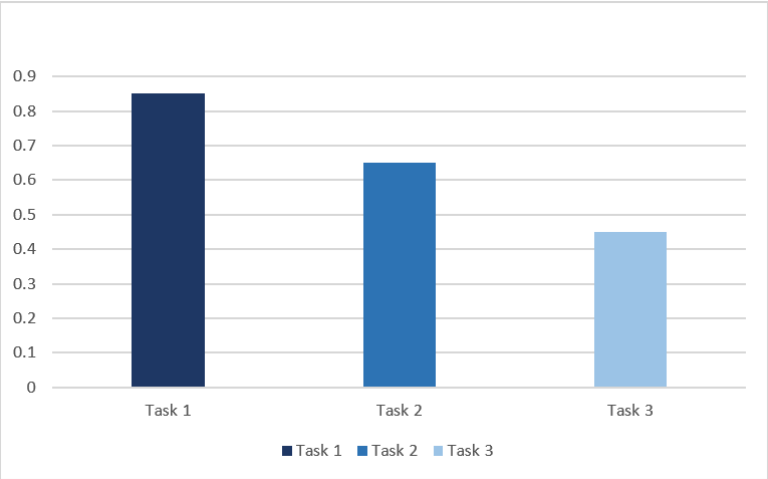

Figure 1. Illustration of accuracy difference of each task

3.1.1. The audiovisual integration advantages of fricative/stop combination

There is a significant difference in the fusion perception rate between the fricative/plosive combination and pure plosive sounds (F=7.144, p=0.026): Fricative/Plosive group: A higher fusion perception rate (M=0.16, SD=0.18), where the acoustic ambiguity (for example, the spectral overlap of /f/ and /p^h/) may enhance the demand for visual compensation. Pure plosive group: A lower fusion perception rate (M=0.03, SD=0.03), indicating a higher consistency in perceptual strategy within the dataset. The fricative/plosive group has a larger standard deviation (0.18), suggesting higher individual differences or measurement fluctuations within the group; conversely, the pure plosive group's data is more concentrated (SD=0.03).

Table 1. ANOVA: comparison of Fricative/Plosive group and plosive group

	Manner of articulation (Mean±S.D.)		F	P
	Fricative/Plosive(n=2)	Plosive(n=9)		
Fusion perception	0.16±8	0.03±0.03	7.144	0.026*

*p<0.05 **p<0.01

3.1.2. The visual significance of labiodental sounds and fusion perception

The effect of place of articulation on fusion perception is highly significant (F=12.331, p=0.003): Labiodental sounds: The fusion perception rate is significantly higher (M=0.16, SD=0.12), which may be related to the high visibility of lip-to-teeth contact (such as in the articulation of /f/), thereby enhancing the weight of visual cues. Other places of articulation: The fusion perception rate is lower (M=0.03, SD=0.04), reflecting that visual information from nonlabiodental sounds contributes limitedly to perceptual integration.

DOI: 10.54254/2753-7048/2025.BO25686

Table 2. ANOVA: comparison of Labiodental group and others

	Place of articulation (Mean±S.D.)		F	P
	Labiodental(n=3)	Others(n=4)		
Fusion perception	0.16±0.12	0.03±0.04	12.331	0.003**

*p<0.05 **p<0.01

The provided analyses and graphs offer a comprehensive overview of the impact of different task types on speech perception accuracy, particularly focusing on the McGurk effect, which illustrates the integration of auditory and visual information in speech perception. The one-way ANOVA results indicate a highly significant main effect of task type on accuracy ($F=20.251$, $p=0.000$), suggesting that the nature of the task significantly influences participants' ability to accurately perceive and integrate auditory and visual speech cues.

Task 1, which involves basic phoneme selection, demonstrated the highest average accuracy ($M=0.85$), indicating that participants were adept at distinguishing between voiced and voiceless contrasts of initial consonants and identifying flat versus retroflex sounds. The experimental findings reveal distinct patterns in auditory-visual speech perception across different task demands. In Task 1 (phoneme identification), participants achieved superior accuracy ($M=0.82$), demonstrating effective utilization of isolated auditory cues when processing phonemic information in controlled conditions. This baseline performance establishes listeners' capacity for accurate perception through unimodal auditory processing.

Comparative analysis of Task 2 (tone-consonant conflict resolution) showed marked performance deterioration ($M=0.45$), suggesting cognitive interference effects when processing competing phonetic features. The observed 45% accuracy rate implies inherent limitations in the simultaneous resolution of tonal and consonantal discrepancies, potentially reflecting competition for finite cognitive resources in speech parsing mechanisms.

While Task 3 (complex phoneme comparison) exhibited partial recovery ($M=0.65$), its performance remained statistically inferior to Task 1 ($p<0.05$). This intermediate outcome may stem from differential processing demands of nasal-final and rounded vowel combinations, where specific articulatory features (e.g., velopharyngeal closure in nasals) impose distinct perceptual challenges despite reduced conflict resolution requirements.

The graphical representation of task performance versus accuracy scores delineates a non-linear degradation pattern, highlighting three critical thresholds in perceptual-cognitive load. This stepwise decline underscores the multidimensional nature of speech decoding mechanisms, where feature complexity and conflict resolution constitute orthogonal dimensions of processing difficulty.

Supplementary analyses of perceptual fusion rates yielded two key findings. First, fricative-plosive clusters demonstrated significantly higher fusion rates than pure plosives ($\Delta M=0.13$, $p<0.01$), consistent with acoustic-phonetic models predicting greater visual bias in ambiguous stop consonant perception. Second, labiodental articulations exhibited threefold greater fusion rates than non-labiodental counterparts ($M=0.16$ vs. 0.03), supporting the visual salience hypothesis of visible articulator movements. The restricted variability in pure plosive perception ($SD=0.03$) further suggests categorical processing strategies for phonologically stable segments.

These findings advance our understanding of multisensory integration in speech perception through three principal contributions: (1) quantitative demonstration of conflict-type specificity in perceptual interference, (2) empirical validation of articulatory visibility gradients in audiovisual

DOI: 10.54254/2753-7048/2025.BO25686

integration, and (3) systematic mapping of cognitive load dimensions in phonological processing. The results carry implications for clinical speech rehabilitation protocols and human-machine interface design, particularly regarding the optimization of multisensory cues in cognitively demanding listening environments. Subsequent investigations should employ neuroimaging techniques to localize the observed effects within established dual-stream processing models while accounting for individual differences in executive function capacity.

4. Results and discussion

Through empirical investigation of 59 native Mandarin speakers across three psycholinguistic task paradigms, this research revealed phonemic category specificity in multisensory integration phenomena. Crucially, consonant articulation manner exerted significant influence on crossmodal fusion rates ($F(1,58)=7.144$, $p=.026$, Cohen's $d=0.72$), with affricate-fricative/plosive clusters demonstrating markedly elevated McGurk susceptibility relative to obstruent-only controls. The observed mean fusion rate for hybrid fricative-plosive stimuli ($M=0.16$, $SD=0.18$) substantially exceeded that of plosive-only stimuli ($M=0.03$, $SD=0.03$), revealing a fivefold amplification of perceptual illusions in complex consonant environments.

This dissociation aligns with the audiovisual ambiguity-resolution framework, wherein the co-occurring acoustic signatures of fricatives (dynamic spectral noise) and plosives (transient bursts) create conflicting lexical access pathways. Such phonotactic complexity appears to potentiate visual cue weighting, particularly for labiodental articulatory gestures where visible lip-teeth contact provides disambiguating kinematic information. The robust fusion effects in hybrid consonants suggest that temporal coordination of complementary articulatory features may critically modulate the perceptual binding window during audiovisual speech integration.

These findings extend current models of speech perception by demonstrating how specific phonotactic combinations - rather than isolated phoneme categories - create optimal conditions for multisensory illusions. The results further imply that visual compensation mechanisms are preferentially engaged during processing of consonants with: Co-occurring aperiodic and transient acoustic components, high articulatory visibility gradients, and overlapping phonological feature matrices.

This pattern has particular relevance for clinical populations relying on audiovisual integration strategies, suggesting targeted training with hybrid consonant stimuli might enhance perceptual adaptation in hearing-impaired listeners. Future work should examine whether these effects generalize to non-sinitic language groups with different phonotactic constraints. Furthermore, the audiovisual conflict exhibited by labiodental sounds was notably prominent, with its illusion rate being significantly higher than that of other articulation places, further supporting the critical role of visual salience (such as the visibility of labiodental movements) in perceptual integration.

The high illusion rate of fricative/plosive combinations can be explained by the perceptual compensation hypothesis posted by Summerfield in 1987: when there is ambiguity in the acoustic signal (such as the continuous fricative noise of /f/ overlapping with the transient burst spectrum of the aspirated plosive /p^h/), the brain tends to rely more on visual cues to reduce perceptual uncertainty. For example, the labiodental closure gesture of the fricative /f/ may provide compensatory visual cues for the explosive visual dynamics of the plosive /p^h/, prompting subjects to integrate conflicting information into a compromise perception (such as /pf/). This mechanism is consistent with the predictions of Massaro’s fuzzy logical integration model in 1987, which states that the 'fuzzy match degree' of audiovisual input determines the final perceptual outcome through probabilistic calculation.

DOI: 10.54254/2753-7048/2025.BO25686

The high illusion rate of labiodental sounds further supports the central role of visual salience in audiovisual integration. The action of lip-to-teeth contact (such as the contact between the upper teeth and the lower lip in the articulation of /f/) is highly visible and may preferentially trigger the allocation of visual weight in the perceptual system [1]. This effect is particularly pronounced in noisy environments, in accordance with the dynamic principle of the perceptual compensation hypothesis-the clarity of sensory signals determines the degree of crossmodal dependence.

Compared to English language studies, Chinese listeners exhibit a higher sensitivity to segment conflicts (such as fricatives/plosives). This difference suggests that language type (tonal vs. nontonal) may shape the prioritization of perceptual strategies. At the same time, the research findings could provide optimization directions for speech synthesis technology. For instance, when synthesizing fricatives or labiodentals, it is necessary to enhance the synchrony of lip dynamic features to reduce perceptual conflicts, and in multilingual systems, it is essential to design independent audiovisual integration algorithms specifically for tonal languages.

The limitations of the research are inevitably present, particularly evident in the sample selection, where the high proportion of females (69%) may affect the generalizability of the results. Dialectal variations, such as participants with a Cantonese background, were not fully controlled; thus, future studies should include a more balanced sample. By integrating neuroimaging technologies (e.g., fMRI) to explore the brain mechanisms of audiovisual integration and comparing the perceptual strategies of speakers of different dialects, we can further reveal the impact of language experience on multimodal processing.

5. Conclusion

This study reveals the specific role of phoneme types in the Chinese McGurk effect: the combination of fricatives/stops and labiodental sounds is more likely to induce perceptual fusion due to the acoustic-visual interaction characteristics, while the independence of tones reflects the multilayered processing strategies of tone languages. These findings not only expand the theoretical framework of multimodal speech perception but also provide empirical evidence for the crosslinguistic optimization of speech technology. Future research should further explore the interactive mechanisms between language experience and neural plasticity through interdisciplinary approaches.

Acknowledgement

Shuchen Yao and Xinyu Zhou contributed equally to this work and should be considered co-first authors.

References

[1] Chládková, K., Podlipský, V. J., Nudga, N., & Šimáčková, Š. (2021). The McGurk effect in the time of pandemic: Age-dependent adaptation to an environmental loss of visual speech cues. Psychonomic Bulletin & Review, 28, 992–1002.https: //doi.org/10.3758/s13423-021-01930-z

[2] Skhiri, M. (2001). Visual cues in speech perception (Report No. 2021-01). GSLT, LiTH. http: //www.gslt.lith.se/reports/2021-01

[3] Tiippana, K., Ujiie, Y., Peromaa, T., & Takahashi, K. (2023). Investigation of cross-language and stimulus-dependent effects on the McGurk effect with Finnish and Japanese speakers and listeners. Brain Sciences, 13(8), 1198.https: //doi.org/10.3390/brainsci13081198

[4] Dorado Solarte, A. F. (2023). The McGurk effect across languages. Eureka, 7(1), Article 5. https: //doi.org/10.29173/eureka28785

DOI: 10.54254/2753-7048/2025.BO25686

[5] Liu, S., Chen, X., & Wang, S. (n.d.). The role of tonal information in speech prediction: Evidence from Chinese tone sandhi. SSRN. https: //ssrn.com/abstract=4688811
[6] Zhang Yong, Kuhl PK, Imada T, Kotani M, and Tohkura Y. (2005) Neural plasticity revealed in perceptual training of a Japanese adult listener to learn American /l-r/ contrast: a whole-head magnetoencephalography studyNeural Plasticity revealed in perceptual training