A preliminary observation on the effect of audio-visual speech perception in voiced cognates by Cypriot-Greek learners of English

Elena Kkese, Dimitra Dimitriou[1]

Cyprus University of Technology, Limassol, [1]UCLan Cyprus, Pyla, Cyprus

ABSTRACT

This study examines the use of audiovisual cues in the perception of sound contrasts which have a different phonemic status in the listeners' L1 and L2. Voice-voiceless cognates differing in the distinctiveness of their visual gestures (/p/-/b/, /t/-/d/, and /k/-/g/) were presented to CG (Cypriot-Greek) learners of English in audio, visual, and audiovisual modalities. Overall identification rates were significantly higher audiovisually in the cases where the auditory and visual information matched (bimodal congruent condition) than in the audiovisual condition in which auditory and visual information did not match (bimodal incongruent condition) or in the audio or video alone condition for either contrast. The results point to the multisensory speech-specific mode of perception, which plays an important role in alleviating the majority of moderate-to-severe L2 difficulties. CG listeners' success seems to depend upon the ability to relate what they see to what they hear.

Access this article online

Website: https://journals.lww.com/jolv/

DOI: 10.4103/jlv.JLV_10_23

Quick Response Code:

Keywords: Audio-visual speech perception, Cypriot-Greek, second language, voiced cognates

INTRODUCTION

L2 (second language) auditory speech perception has been an area that has long attracted interest due to several perceptual difficulties encountered by L2 learners. In such conditions, in which speech is degraded, the addition of visual cues may greatly benefit L2 listeners; it seems that there is a direct relationship between auditory and visual characteristics of speech.[1-3] This can be attributed to both segmental and suprasegmental cues; segmental information can disambiguate specific distinctions in terms of place and manner of articulation, whereas suprasegmental cues provided by the visual channel may be better used.

L2 speakers, just like L1 (first language) speakers, rely heavily on the acoustic signal during speech perception, although developing an acute awareness of acoustic distinctions in an L2 takes time and requires exposure to the L2 sounds.[4,5] If their acoustic understanding of the L2 distinctions is not adequately developed, then L2 learners' reliance on the acoustic signal may lead to inaccurate perception. Therefore, segmental perception is particularly challenging for L2 learners, especially when the inventories of the L1 and L2 contain different phonemes or allophones as in the case of Cypriot-Greek (CG) compared to Standard Southern British English (SSBE), which is the variety used across the world and in previous studies.[6,7] Specifically, CG involves 51 consonants, these being /p pʰ: b t tʰ: d c cʰ: ɟ k kʰ: g f f: v v: θ θ: ð ð: s s: z z: ʃ ʃ: ʒ ʒ: ç ç: j j: x x: ɣ ɣ: ts ʧ ʧ: ʤ m m: n n: ɲ ŋ l l: ʎ ɾ r/; on the contrary, SSBE involves 24 consonants /p b t d k g f v θ ð s z ʃ ʒ h ʧ ʤ m n ŋ ɹ j w l/.[7] Nonetheless, CG is much more transparent than SSBE in the representation of phonology since most graphemes in the language represent a single phoneme.[8]

Address for correspondence:
Dr. Elena Kkese,
Tasou Isaak 3, Alethriko, Larnaca 7570, Cyprus.
E-mail: elenakkese@hotmail.com

How to cite this article: Kkese E, Dimitriou D. A preliminary observation on the effect of audio-visual speech perception in voiced cognates by Cypriot-Greek learners of English. J Laryngol Voice 2023;13:30-5.

Received: 23-May-2023; **Accepted:** 25-Sep-2023; **Published:** 13-Jun-2024.

This study aimed to investigate the effectiveness of auditory and audiovisual cues for L2 perception for contrasts in which the visual cues vary in terms of informativeness for L2 listeners. The study seems to be the first to investigate the audiovisual perception of the distinction between English /p b t d k g/ in CG L2 learners of English as in CG, some descriptions suggest that the voiced plosives /b d g/ are absent or are realized as prenasalized voiced plosives as in [mba'mbas].[7,9-12] Previous research comparing L1 CG and L2 English refers to auditory perception studies including voiced-voiceless cognates' perception on a word or utterance level,[7,13,14] as well as consonant and vowel perception on a word level.[15,16] Therefore, the purpose of this study was to examine the perception of CG learners of L2 English, an attempt to investigate to what extent participants depend on the auditory and/or visual cues in the specific context given that multi-modal perception is expected to yield speech perception.

Research questions

The study attempts to answer the following questions:

1. What is the effect of visual cues in enhancing the perception of certain types of phonetic information such as voicing and place of articulation, particularly for plosive consonants?
2. To what extent combined audiovisual cues induce effects larger than those elicited by either cue on its own?

The research approach used for the current study is quantitative aiming at identifying the difficult phones in L2 English with reference to plosive consonants in the different modes of perception. Differences are examined in the dependent variable (percentage of correctness) thought to be caused by the independent variables (different categories).

MATERIALS AND METHODS

Ethics

The study followed the protocol of the Helsinki Declaration of 1975. Ethical approval was obtained from the institutional research and ethics committee, and consent was obtained from each participant.

Study design

Selection and description of participants

Fourteen CG users of L2 English with normal hearing and vision were recruited for the purpose of this study. These involved seven males and seven females between the ages of 18 and 26, who were active undergraduate university students. Before they were asked to undertake the developed perceptual tasks, the participants had to fill in a consent form and a brief background questionnaire to ensure that they shared the same characteristics. These referred to age, gender, L1, educational level, and exposure to CG as the aim was to eliminate inter-group differences. Participation in the task was on a completely voluntary basis and students were ensured about the confidentiality of their personal details.

The research period involved one fall semester with data being collected in one session. Participants had to take the task on a computer in a quiet testing room with audio presented over earphones set to a comfortable volume (75 dB). They had to complete the perceptual tasks in which four conditions were created for plosive perception: auditory-only, video-only, audiovisual congruent, and audiovisual incongruent. During the first part, participants were asked to take the two unimodal tasks of the study; this was followed by the two bimodal tasks in the second part of the study. For the first part, participants had to report what they could perceive in the auditory-only condition and what was being said while watching the lips on the screen in the visual-only condition (without sound). A short 2-min break followed between the two short unimodal tasks. After the unimodal part, there was a short pause halfway (5 min); after that, participants completed the bimodal part of the perceptual task. Participants for the two conditions had to report what they could perceive in the audiovisual condition. However, for the congruent bimodal stimuli, auditory and visual information matched while for the incongruent bimodal stimuli, the auditory and visual information did not match. Participants could also have a 2-min break between the bimodal part if they wanted. They could control the transition between the slides; they had to indicate whether the pseudowords began by /p b t d k g/ by pressing the corresponding button on a response pad. All tasks were built using PsychoPy. The speaker was a female adult, who was a native speaker of CG; she was recorded from her shoulders up and was instructed to speak naturally in an emotional passive tone without moving her head. She was recorded at a 44.1 kHz sample while speaking into a microphone that fed directly into the sound card (IDT High-Definition Audio CODEC) of a laptop computer. Overall, testing lasted about 15 min while no feedback was provided in the four conditions, and the target stimuli were not repeated.

Materials

Unimodal and bimodal perceptual tasks

For this study, four conditions were created; these consisted of the auditory- and visual-only stimuli constituting the unimodal part of the perceptual task and the congruent and incongruent audiovisual stimuli constituting the bimodal part of the study. For all conditions, the same 72 disyllabic pseudowords of **C**VCV structure were arranged in 12 minimal sets consisting of six items each; the target consonant was found word-initially; vowel was one of the following: /a e o/. Three minimal sets were included for each category, namely for the voiceless/voiced bilabial /p b/, the alveolar /t d/,

and the velar /k g/, making up nine minimal sets in total to enable a comparison of the three plosive categories as well as of the voiced-voiceless cognates. Distractors focusing on the voicing contrast were also intermixed and made up three of the minimal set words; these included fricative consonants such as the labiodental [f] and [v], dental [θ] and [ð], and alveolar [s] and [z]. PsychoPy was used to create the experimental presentation of the edited stimuli and collect response and reaction time data. A PowerPoint presentation of 12 minimal sets for each condition was created that was presented using a Dell computer. Overall, responses were scored as correct or incorrect generating an overall percent correct score. Similarly, percent correct scores for each consonant category were also obtained.

RESULTS

Figure 1 shows the overall performance of participants in identifying plosives in each condition ($N = 14$). A one-way repeated measures analysis of variance (ANOVA) was conducted to assess participants' performance across the four conditions (Unimodal Audio, Unimodal Video, Bimodal Congruent and Bimodal Incongruent). CONDITION was set as the within-subject factor with four levels and a significant effect was found ($P < .001$, Wilks' Lambda = .026, $F = 137.811$, $\eta 2 = .974$). Pairwise comparisons showed significant differences between Bimodal Congruent and each of the other conditions with $P < .001$ in all comparisons. A significant difference was also found between Unimodal Audio and Unimodal Video ($P = .020$).

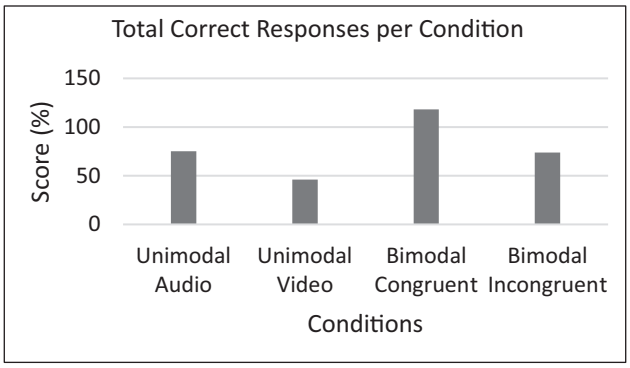

Figure 1: Correct and incorrect responses to stimuli in each condition

Figure 2 shows participants' performance in identifying stimuli with word-initial voiced or voiceless plosives in each condition. A two-way repeated measures ANOVA was conducted to evaluate the influence of CONDITION, VOICING, and the CONDITION*VOICING interaction on participants' performance. CONDITION with four levels and VOICING with two levels (Voiced and Voiceless) were set as the within-subject factors. A significant main effect of CONDITION, VOICING and CONDITION*VOICING was observed (CONDITION: $P < .001$, $F = 38.101$; VOICING: $P < .001$, $F = 62.52$; CONDITION*VOICING: $P < .001$, $F = 6.674$). Pairwise comparisons showed significant differences between Voiced and Voiceless plosive scores overall, with $P < .001$. Significant differences in the scores of participants between voiced and voiceless plosives were observed in all conditions except for Unimodal Video (Unimodal Audio: $P = .009$; Bimodal Congruent and Bimodal Incongruent: $P < .001$).

Figure 3 shows participants' performance in identifying target bilabial, alveolar, and velar plosives in each of the four conditions. A two-way repeated measures ANOVA with CONDITION and PLACE as the within-subject factors with four and three levels respectively (Bilabial, Alveolar, Velar for PLACE) was conducted to evaluate their influence on participants' scores. A significant main effect

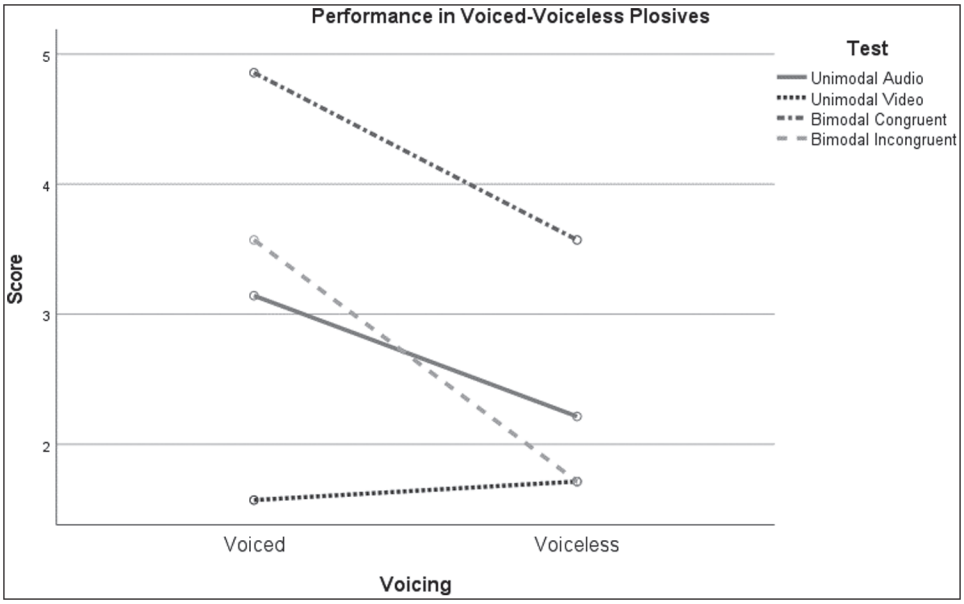

Figure 2: Mean score of participants in target voiced or target voiceless plosives in each condition

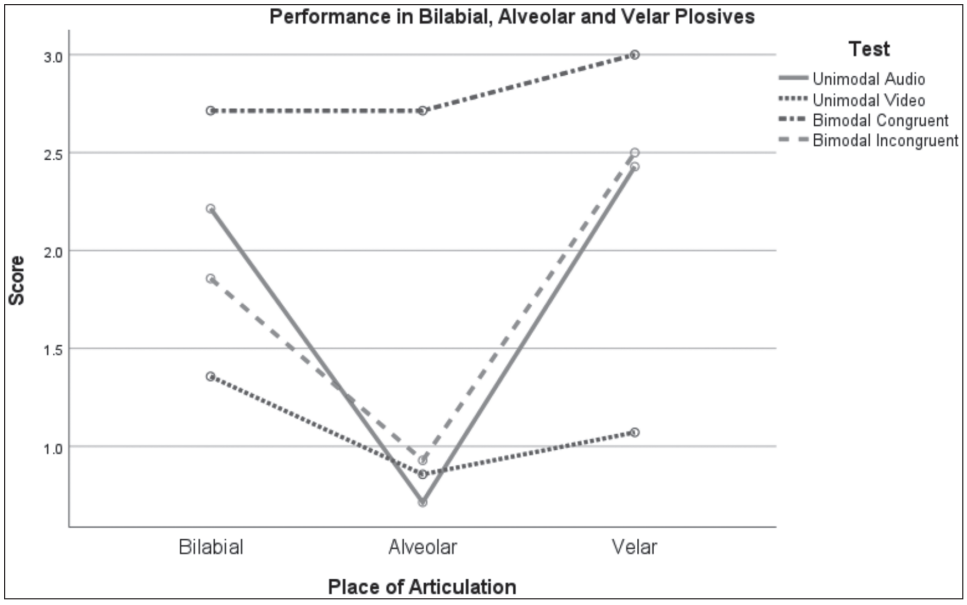

Figure 3: Mean score of participants in target bilabial, alveolar, and velar plosives in each condition

of CONDITION ($P < .001$, $F = 38.101$), PLACE ($P < .001$, $F = 41.777$) and CONDITION*PLACE ($P < .001$, $F = 6.381$) was observed. Pairwise comparisons showed significant differences between Bilabial and Alveolar plosives and between Alveolar and Velar plosives overall ($P < .001$ in both cases). In the Unimodal Audio condition, differences reached significance between Bilabial and Alveolar plosives, and between Alveolar and Velar plosives ($P < .001$ in both cases). The same pattern of identification of plosives was observed in the Bimodal Incongruent condition: Bilabial-Alveolar and Alveolar-Velar plosive identification differed significantly ($P = .001$ and $P < .001$, respectively). In the Unimodal Video and Bimodal Congruent conditions, no significant differences were observed based on the place of articulation of the plosive.

The availability of stimuli with fricative consonants (distractors) enabled a comparison of participants' performance in target fricatives and plosives. This comparison could reveal further tendencies in these learners' perceptions of L2 consonants in different conditions, given that the focus of this study was on plosive consonants. Figure 4 shows the percentage of correct responses to stimuli with initial plosive and stimuli with initial fricative consonants, illustrating a tendency for fricative consonants to be better identified by participants in all conditions. Furthermore, participants' identification of fricative consonants in each condition mirrors their performance in plosive consonant identification. However, given the limited number of data collected for fricative consonants and the imbalance in data collected for plosive compared to fricative consonants, no further statistical analyses were conducted to ascertain whether this difference is statistically significant.

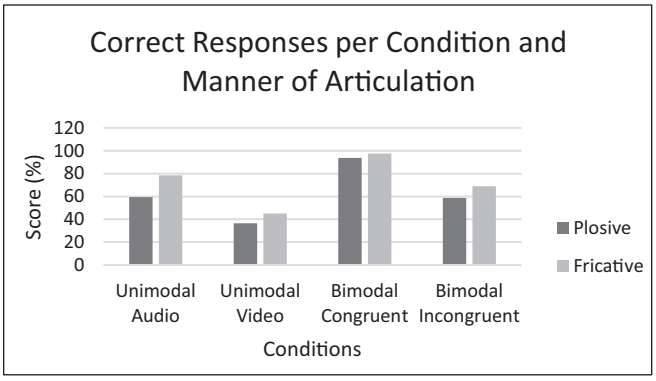

Figure 4: Correct and incorrect responses to distractors in each condition

DISCUSSION

This study aimed to assess the role of visual information in L2 speech acquisition, and more specifically, in the perception of plosive consonants by CG learners of English. The study emphasized on plosive consonants, as they are considered to be difficult for these learners, given the differences between the L1 and L2 consonantal inventory.[14] More specifically, according to some descriptions, voiced plosives do not exist in CG or are realized as prenasalized voiced plosives.[7,12-16]

In relation to the first research question, this study suggests that visual information can aid the perception of certain types of phonetic information, i.e. voicing and place of articulation. Given that CG contains voiceless plosives, better identification was expected in this category. However, participants' identification accuracy was higher for voiced compared to voiceless plosives in all conditions, except in the video-only condition, suggesting that voiced

plosives were not substituted with the closest L1 sounds in place and manner of articulation, which participants could perceive based on L1 phonetic inventory knowledge.[17-19] In terms of place of articulation, participants performed better in identifying bilabial and velar compared to alveolar plosives, especially in the Unimodal Audio and Bimodal Incongruent conditions. This was expected for velar plosives, given that they are produced with longer VOT values,[20-22] but not for bilabial plosives, which are produced with the shortest VOT values.[22,23] These findings suggest that visual cues can enhance the perception of certain types of phonetic information such as voicing and place of articulation and can play an important role in the perception of plosive consonants. In relation to manner of articulation, this study focused on plosive consonants, since the acoustic difference between sounds with a different manner of articulation, for example, /b/-/v/, may be more easily perceived, especially since they are also marked by visual cues. However, given the availability of the three sets of distractors involving fricative consonants, a descriptive analysis was conducted to identify possible differences in the identification of consonants with a different manner of articulation. This analysis, albeit restricted by the small number of data, shows that participants' performance in the identification of fricative consonants follows the same pattern as in plosives.

Turning to the second research question, the results show that the combination of audiovisual cues facilitates perceptual performance, compared to conditions where each cue is provided on its own. This is evident from participants' near-ceiling performance when provided with auditory information that matched visual information (Bimodal Congruent task), compared to all other conditions, in which participants performed significantly worse. The influence of the likelihood of an integrated multisensory percept has been examined in several studies, by studying the timing of single auditory and visual events[24,25] or simple periodic modulations of stimulus features[24,26,27]; however, this is different compared to natural sounds, such as speech. A study[28] addressed this issue using timed sequences of discrete auditory-visual events and found that coherence discrimination was better for unpredictable compared to predictable sequences.

During speech perception, both L1 and L2 listeners rely primarily on the acoustic signal. However, an acute awareness of acoustic distinction requires exposure to the L2 sounds, and when the acoustic understanding of L2 distinctions is not adequately developed, as in the case of CG learners of English, the acoustic signal may not provide sufficient information for accurate perception, meaning that relying on a single cue alone may be challenging. Therefore, incorporating visual cues may facilitate perception, as suggested by the results of this study. The integration of auditory and visual signals seems to benefit L2 speech perception and points to a strong McGurk effect. This study shows the multisensory speech-specific mode of perception, which is important in addressing and minimizing L2 difficulties. An important implication of this is that CG learners could benefit from explicit instruction of visual and auditory distinctions, given that their perceptual performance seems to depend on whether they can relate what they see to what they hear. This can be better understood when the CG orthography is taken into account, which employs a more transparent system (one-to-one correspondence between graphemes and phonemes) in contrast to SSBE that uses a more opaque system (one-to-many correspondences between graphemes and phonemes). Therefore, CG learners could benefit from being exposed to the multisensory speech-specific mode of perception learning to pay more attention to auditory cues.

Financial support and sponsorship
Nil.

Conflicts of interest
There are no conflicts of interest.

REFERENCES

1. Cappelletta L, Harte N. Phoneme-to-viseme Mapping for Visual Speech Recognition. Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods 2012;2:322-9.
2. Massaro DW, Cohen MM, Gesi AT. Long-term training, transfer, and retention in learning to lipread. Percept Psychophys 1993;53:549-62.
3. Potamianos G, Neti C, Gravier G, Garg A, Senior AW. Recent advances in the automatic recognition of audio-visual speech. Proc IEEE 2003;91:1306-26.
4. Flege JE, Liu S. The effect of experience on adults' acquisition of a second language. Stud Second Lang Acquis 2001;23:527-52.
5. Flege JE. Give input a chance!. In: Piske T, Young-Scholten M, editors. Input Matters in SLA. Multilingual Matters; 2009. pp. 175–90.
6. Iverson P, Kuhl PK, Akahane-Yamada R, Diesch E, Tohkura YI, Kettermann A, *et al*. A perceptual interference account of acquisition difficulties for non-native phonemes. Cognition 2003;87:B47-57.
7. Kkese E. Identifying plosives in L2 English: The case of L1 Cypriot Greek speakers. Bern: Peter Lang; 2016.
8. Kkese, E. L2 writing assessment: The neglected skill of spelling. Cambridge Scholars Publishing; 2020.
9. Arvaniti A. Linguistic practices in Cyprus and the emergence of Cypriot Standard Greek. Mediterr Lang Rev 2006;17:15-45.
10. Kappa I. On the Acquisition of Syllable Structure in Greek. J Greek Linguist 2002;3:1-52.
11. Terkourafi M. Politeness in Cypriot Greek: A Frame-Based Approach. (Ph.D. dissertation). Cambridge, England: University of Cambridge; 2001.
12. Newton B. The Generative Interpretation of a Dialect. A Study of Modern Greek Phonology. Cambridge: Cambridge University Press; 1972.
13. Kkese E, Petinou K. Perception Abilities of L1 Cypriot Greek Listeners – Types of Errors involving Plosive Consonants in L2 English. J Psycholinguist Res 2017a;46:1-25.
14. Kkese E, Petinou K. Factors affecting the perception of plosives in second language English by Cypriot-Greek listeners. In: Babatsouli E, editor. Proceedings of the International Symposium on Monolingual and Bilingual Speech 2017. 2017. pp. 162-7.

15. Kkese E, Karpava K. Applying the Native Language Magnet Theory to an L2 setting: Insights into the Cypriot Greek adult perception of L2 English. In: Babatsouli E, editor. Proceedings of the International Symposium on Monolingual and Bilingual Speech 2019. 2019. Chania, Greece: Institute of Monolingual and Bilingual Speech. pp. 67-74.
16. Kkese E, Karpava S. Challenges in the perception of L2 English phonemes by native speakers of Cypriot Greek. J Monoling Biling Speech 2021;3:1-39.
17. Carlisle RS. Markedness and environment as internal constraints on the variability of interlanguage phonology. In: Yavas M, editor. First and Second Language Phonology. San Diego: CA: Singular Publishing Group Inc.; 1994. pp. 223–249.
18. Eckman F. Markedness and the contrastive analysis hypothesis. Lang Learn 1977;27:315-30.
19. Weinreich U. Languages in Contrast: Findings and Problems. The Hague: Mouton; 1953.
20. Ng M, Chen Y, Wong S, Xue S. Interarticulator timing control during inspiratory phonation. J Voice 2011;25:319-25.
21. Liu H, Ng M, Wan M, Wang S, Zhang Y. Effects of place of articulation and aspiration on voice onset time in Mandarin esophageal speech. Folia Phoniatr Logop 2007;59:147-54.
22. Lisker L, Abramson AS. A cross-language study of voicing in initial stops: Acoustical measurements. Word 1964;20:384-422.
23. Klatt DH. Vowel lengthening is syntactically determined in a connected discourse. J Phonetics 1975;3:129-40.
24. Fujisaki W, Nishida S. Temporal frequency characteristics of synchrony–asynchrony discrimination of audio-visual signals. Exp Brain Res 2005;166:455-64.
25. Zampini M, Guest S, Shore DI, Spence C. Audio-visual simultaneity judgments. Percept Psychophys 2005;67:531-44.
26. Recanzone GH. Auditory influences on visual temporal rate perception. J Neurophysiol 2003;89:1078-93.
27. Spence C, Squire S. Multisensory integration: Maintaining the perception of synchrony. Curr Biol 2003;13:R519-21.
28. Denison RN, Driver J, Ruff CC. Temporal structure and complexity affect audio-visual correspondence detection. Front Psychol 2013;3:619.