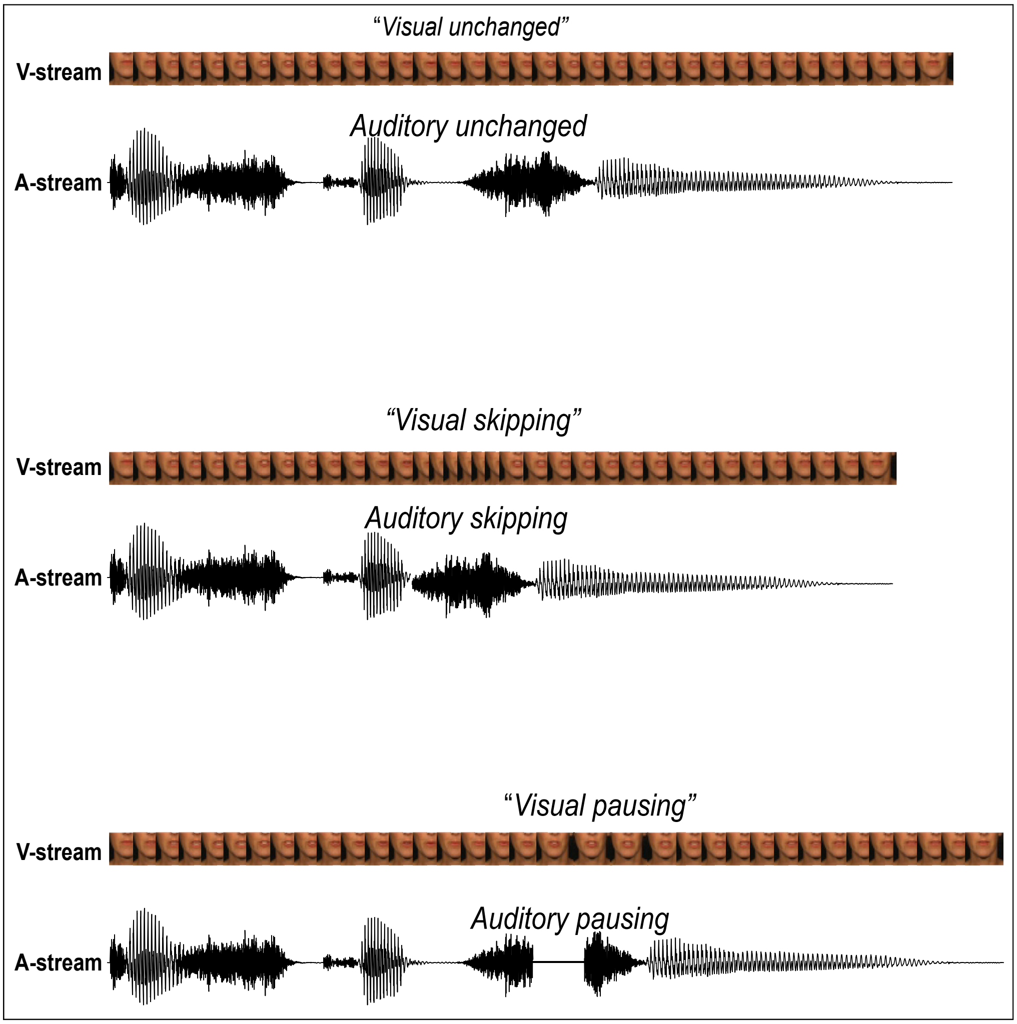
iScience

CellPress
OPEN ACCESS

Article

Audition controls the flow of visual time during multisensory perception



Mariel G. Gonzales, Kristina C. Backer, Yueqi Yan, Lee M. Miller, Heather Bortfeld, Antoine J. Shahin

ashahin@ucmerced.edu

Highlights

We describe the significance of the Audiovisual Time-Flow Illusion

Temporal perturbations to auditory speech drive perception of visual speech

However, perturbing visual speech stimuli does not affect auditory perception

Auditory processing controls the temporal perception of the visual speech stream

Gonzales et al., iScience *25*, 104671
July 15, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.104671

iScience

CellPress
OPEN ACCESS

Article

Audition controls the flow of visual time during multisensory perception

Mariel G. Gonzales,[1] Kristina C. Backer,[1,2] Yueqi Yan,[2] Lee M. Miller,[3,4,5] Heather Bortfeld,[1,2,6] and Antoine J. Shahin[1,2,7,*]

SUMMARY

Previous work addressing the influence of audition on visual perception has mainly been assessed using non-speech stimuli. Herein, we introduce the Audiovisual Time-Flow Illusion in spoken language, underscoring the role of audition in multisensory processing. When brief pauses were inserted into or brief portions were removed from an acoustic speech stream, individuals perceived the corresponding visual speech as *"pausing"* or *"skipping*", respectively—even though the visual stimulus was intact. When the stimulus manipulation was reversed—brief pauses were inserted into, or brief portions were removed from the visual speech stream—individuals failed to perceive the illusion in the corresponding intact auditory stream. Our findings demonstrate that in the context of spoken language, people continually realign the pace of their visual perception based on that of the auditory input. In short, the auditory modality sets the pace of the visual modality during audiovisual speech processing.

INTRODUCTION

The visual modality's influence on auditory perception of spoken language is well established. A classic example of this is the McGurk illusion, wherein the incongruent pairing of visual and auditory speech results in the auditory perception of either the visually conveyed phoneme or a different, third phoneme (McGurk and MacDonald, 1976; Abbott and Shahin, 2018; Shahin et al., 2018). However, the impact of the auditory modality on visual perception in spoken language processing is less understood. Examples from the non-speech domain (Welch and Warren, 1980; Recanzone, 2003; Stein et al., 1996; Sekuler et al., 1997; Shams et al., 2000, 2001, 2002; Vroomen and de Gelder, 2000, 2004; Chen and Vroomen, 2013) suggest a critical role for audition in shaping visual perception. For example, in the Double Flash Illusion (Shams et al., 2000, 2001, 2002) when one flash is paired with two tones, individuals often see two flashes. However, when two flashes are paired with one tone, individuals often see one flash. In the flash-lag effect (FLE), perception of a flash of light lags relative to a moving object depending on when bursts of sound are played. The FLE was found to vary linearly with when the bursts of sound occur, that is, either before, during, or after each flash (Vroomen and de Gelder, 2004). These findings suggest that in audiovisual crossmodal perception, the auditory modality is used as a temporal reference for the visual modality. However, the above-mentioned illusions cannot address two crucial questions: i) whether audition exerts such an influence over vision in the context of complex, real world multisensory stimuli such as audiovisual speech, and ii) whether audition merely serves as a temporal reference for vision, or whether it actually controls the *flow of perceived visual time* to maintain crossmodal alignment.

We posited that the role of the auditory modality is to set the pace of the visual modality by controlling (interrupting or modulating) perceived visual time, so that visual temporal processing synchronizes with the information conveyed by audition. In the experimental blocks, individuals attended to videos of a speaker uttering trisyllabic words. In one condition, the acoustic speech signal had two silent segments inserted (*A-pause*), inducing a perception that the acoustic speech was "*pausing*" or slowing-down; in another condition, the audio had two speech segments excised (*A-skip*), inducing a perception that the acoustic speech was "*skipping*" or speeding-up. Figure 1 depicts example waveforms and spectrograms of manipulated audios. (Note that we use quotes to indicate subjective perception, whereas an absence of quotes denotes physical stimulus properties.) The visual part of the video remained unchanged. The experimental blocks also included intact audiovisual stimuli (*unchanged*) and *catch* trials whereby both the audio and visual parts of the videos were *pausing* (*AV-pause*) or *skipping* (*AV-skip*) (i.e., temporally aligned segments were inserted or excised across both modalities). Individuals made judgments about whether the visual part of the video was *"pausing"* (*"V-pause"*), *"skipping"*

[1]Department of Cognitive and Information Sciences, University of California, 5200 N Lake Rd, Merced, CA 95343, USA

[2]Health Sciences Research Institute, University of California, 5200 N Lake Rd, Merced, CA 95343, USA

[3]Center for Mind and Brain, University of California, Davis, CA 95618, USA

[4]Department of Neurobiology, Physiology and Behavior, University of California, Davis, CA 95616, USA

[5]Department of Otolaryngology / Head & Neck Surgery, University of California, Davis, Sacramento, CA, 95817, USA

[6]Department of Psychological Sciences, University of California, Merced, CA 95343, USA

[7]Lead contact

*Correspondence: ashahin@ucmerced.edu

https://doi.org/10.1016/j.isci.2022.104671

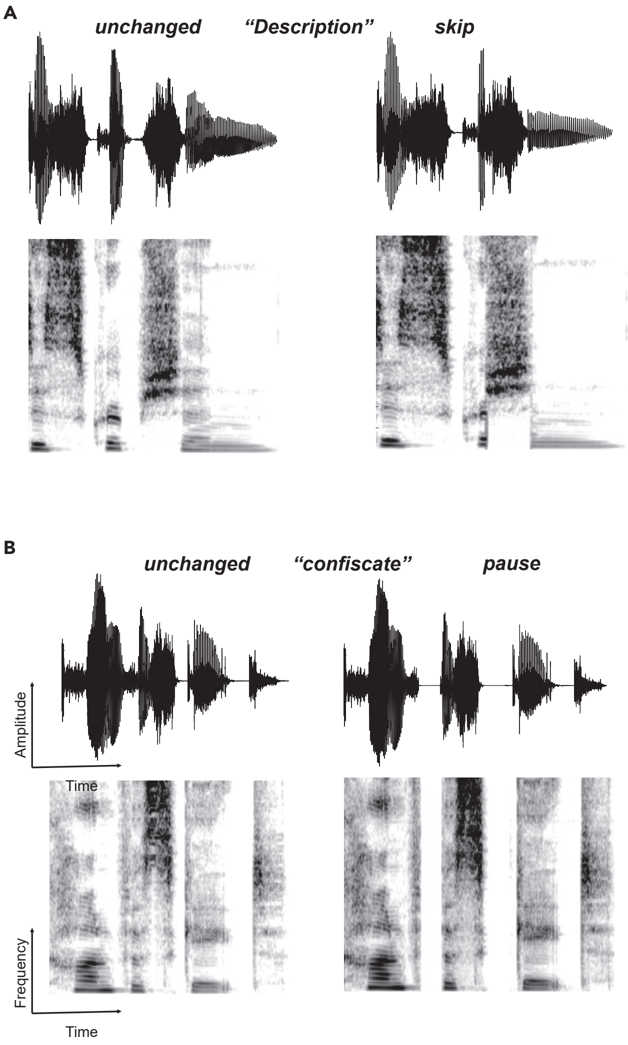


Figure 1. Example audio waveforms and spectrograms

(A) Example audio waveforms of two manipulated words, that led to induction of "skipping" and "pausing" perception in the visual modality.
Example audio spectrograms of the two manipulated words shown in (A).

(*"V-skip"*), or remained *"unchanged"* (*"V-unchanged"*). In the control blocks, the stimulus manipulations were reversed. That is, individuals listened to intact audio streams paired with manipulated visual streams (*V-skip*, *V-pause* or *unchanged*) while making judgments about whether the auditory part was *"pausing"* (*"A-pause"*), *"skipping"* (*"A-skip"*), or remained *"unchanged"* (*"A-unchanged"*). As in the experimental blocks, the control blocks also included intact and *catch* trials. Within each block, the stimuli were randomly intermixed, and the order of the experimental and control blocks was counterbalanced across subjects.

We hypothesized that individuals would perceive more robust "*pausing" or* "*skipping"* in the visual speech (*"V-pause"* or *"V-skip"*) when the auditory speech *paused or skipped (A-pause or A-skip)*, respectively, relative to the opposite manipulation—judging intact acoustic stimuli as "*A-pause"* or "*A-skip,"* when the visual stimuli *paused* or *skipped* (*"V-pause"* or *"V-skip"*), respectively. Such a finding would indicate that the auditory modality sets the pace of the visual modality, and not the other way around.

RESULTS

Figure 2 (see also Table S1: Mean response percentages and standard deviations, relating to Figure 2) depicts group percentages for "*pause"* and "*skip"* illusory percepts, as well as *"unchanged"* percepts. These

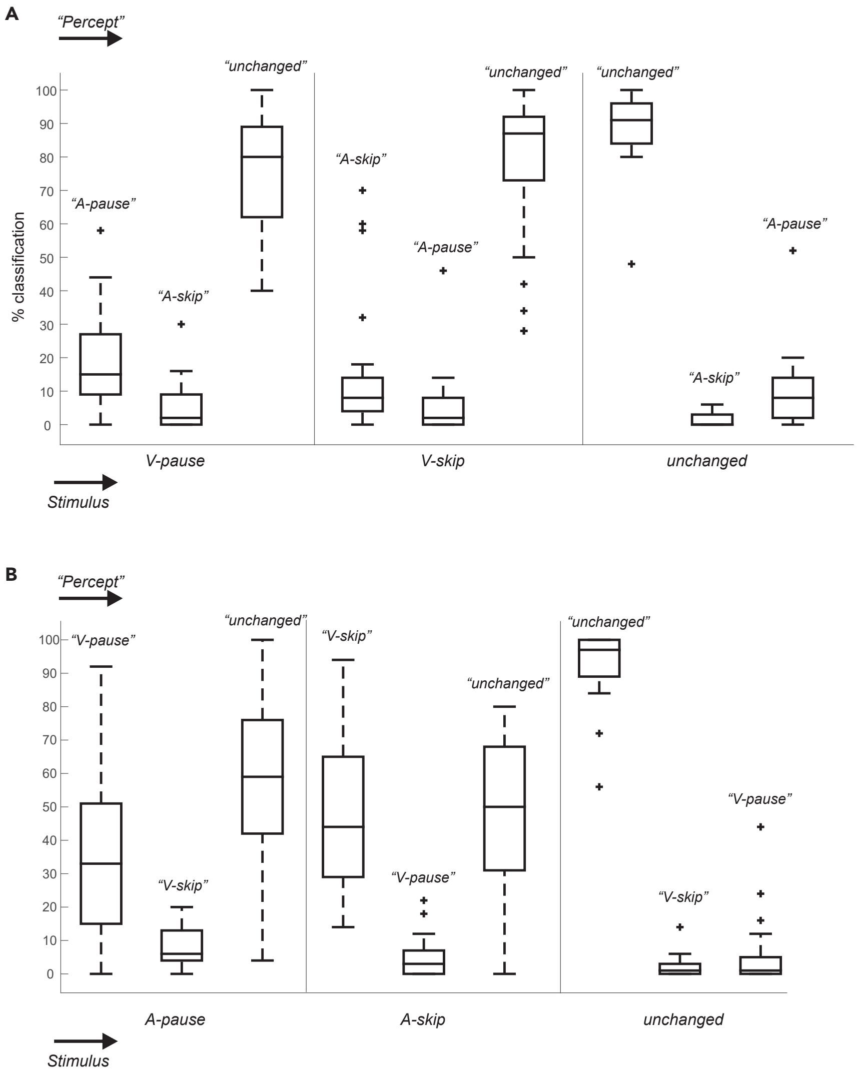


Figure 2. Box plots of response (Percept) percentages in the auditory (A) or visual (B)modality induced by stimulus manipulation in the counter modality

For example, the percentage of the percept *"A-pause"* in response to a stimulus of *V-pause* indicates the percentage that individuals perceived a *"pause"* in the auditory stream in responses to a physical *pause* in the visual stream that is paired with an acoustically intact auditory stream.

percentages were calculated as the number of trials perceived as either "*pause,*" "*skip,*" *or* "*unchanged*" in one modality, divided by all trials within the modality (see examples in Figure 2 and Table S1). In the experimental blocks, the average percentage of individuals who experienced a "*pause*" illusion in the intact visual stream ("*V-pause*") because of a physical *pause* in the auditory stream (*A-pause*) was 35.4%, significantly different from the percentage experiencing a *"pause"* illusion in the *unchanged* trials (5.3%, p < 0.001). Participants perceived a "*skip*" illusion in the intact visual stream ("*V-skip*") because of a physical *skip* in the auditory stream (*A-skip*) 48% of the time, which was also significantly different from the average probability of experiencing a *"skip"* illusion in the *unchanged* trials (2.2%, p < 0.001). In the control blocks, individuals experienced an "*A-pause*" or "*A-skip*" illusion in the intact auditory stream because of a *pause or skip* in the visual stream (*V-pause* or *V-skip*) 18.1 and 14.7% of the time, respectively. Both percentages were significantly different from the same percentages for the *unchanged* trials (*V-pause"* perceived as *"A-pause"* vs. *unchanged* perceived as *"A-pause"*: 18.1 vs. 9.6%, p < 0.001; *V-skip* perceived as *"A-skip"* vs. *unchanged* perceived as *"A-skip"*: 14.7 vs. 1.5%, p < 0.001).

We next conducted mixed effects logistic regression analysis to test our hypothesis directly—whether the illusions observed in the visual stream because of auditory stimulus alterations were more robust than the illusions observed in the auditory stream because of visual stimulus alterations. Results demonstrated a significant difference in the likelihood of individuals experiencing the *"pause"* or *"skip"* illusions in the visual stream when the alterations occurred in the auditory streams, versus experiencing the illusions in the auditory stream when the alterations occurred in the visual streams [for *A-pause* perceived as *"V-pause"* vs. *V-pause* perceived as *"A-pause"*, [*"V-Pause"* vs. *"A-Pause"*, Odds ratio (OR) = 0.36 [0.29, 0.44], p < 0.001; for *"V-Skip"* vs. *"A-Skip"*, OR = 0.17 [0.14, 0.20], p < 0.001].

Finally, we evaluated the differences in *unchanged* trials (no change in either audiovisual stream) occurring in the experimental and controls conditions and *catch* (*AV-pause*, *AV-skip*) trials (temporal manipulation in both modalities) occurring in the experimental and control conditions, to ascertain whether participants across both tasks were performing comparably. No significant change was observed across the main and control blocks for the *catch* trials (for *AV-pause* perceived as *"V-pause"* in main versus *AV-pause* perceived as *"A-pause"* in control, OR: 1.23[0.70, 2.17], p = 0.471; for *AV-skip* perceived as *"V-skip"* in main vs. *AV-skip* perceived as *"A-skip" in* control, OR: 1.65[0.61, 4.47], p = 0.324). However, in the *unchanged* conditions, performance differed significantly between the experimental and control blocks. Despite the means being close to the same ("*V-unchanged*" = 92.5%; "*A-unchanged*" = 89%), participants were more accurate in labeling the intact *unchanged* streams as "*V-unchanged*" (main blocks) than the intact *unchanged* streams as "*A-unchanged*" (control blocks) [for *unchanged* correctly perceived in main vs. control, OR: 0.64[0.48, 0.85], p < 0.001]. In short, participants' comparable performance on the *unchanged* (despite a significant difference) and *catch* trials across the experimental and control blocks demonstrates that participants were vigilant in their performance of the tasks. These results reinforce the findings observed during the unimodal *"skip"* and *"pause"* manipulations, lending additional support to the interpretation that the illusory effects observed between the experimental relative to the control blocks are true perceptual phenomena.

DISCUSSION

Our results show that the auditory modality, not the visual modality, sets the pace of audiovisual speech processing, whereby the visual modality "*skips*" and "*pauses*" to keep pace with the rate of information conveyed by the auditory modality. This is in line with a temporal reference role for the auditory modality in multisensory processing, but goes well beyond it, showing that audition actually controls the flow of time for visually perceived events. In ecologically appropriate situations, sounds often lead us to redirect our visual attention, with the visual modality recalibrating its spatial focus. In other circumstances, e.g., discourse, rather than recalibrating its spatial focus, vision recalibrates its temporal focus as demonstrated here.

The current findings raise the question of *why* the visual modality synchronizes with the pace of the auditory input, but not the reverse. In natural discourse, the visual modality (mouth movement) often precedes corresponding sound production (Chandrasekaran et al., 2009; Schwartz and Savariaux, 2014) and provides predictive cues to optimize the processing of sound. Thus, visual input plays a major role, behaviorally and neurophysiologically, in guiding our predictions about the unfolding speech signal (Besle et al., 2004; van Wassenhove et al., 2005). This predictive quality enhances speech comprehension, especially in noisy situations (Sumby and Pollack, 1954). Here we provide strong complementary evidence that audition, the more temporally precise modality, controls the flow of time in vision, the more temporally predictive modality. This raises important mechanistic and phenomenological questions about neural computational time versus experienced time, not only across sensory modalities but in sensorimotor time as well (De Kock et al., 2021). In view of the current findings, we posit that as the speech stream changes pace, the visual modality realigns its pace according to the auditory modality to reassume a leading position. In this way, the visual modality maintains its predictive impact on audition and optimizes speech intelligibility.

Limitations of the study

Our understanding of the Audiovisual Time-Flow illusion remains limited with regard to the neural and contextual mechanisms that drive it. For example, it is unclear whether the effect is associated with transient changes in the auditory stream as in the present experiment, or if it is also observed during a slow buildup of audiovisual stream misalignments, such as when the audio is sped up or slowed down. A second limitation is whether this phenomenon is speech specific, a question that should be addressed through the use of nonspeech stimuli, such as a musical performance or a movie clip of an inanimate object's movement. A third limitation, building on the previous two, is whether this phenomenon is driven by low-level correspondence between the auditory and visual streams or is linguistically driven. For example, if individuals can still experience the illusion with incongruent speech streams, whereby phonemes and visemes are misaligned, one can conclude that the illusion is driven by low level correspondence of AV streams as opposed to higher level linguistic factors. Finally, it is imperative to assess the neurophysiology that facilitates the transfer of synchrony from the auditory modality to the visual modality to determine whether this transfer is directly communicated between nominally unisensory cortices or mediated by a higher-order multisensory hub.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Participants
- METHOD DETAILS
  - ○ Stimuli *experimental blocks*
  - ○ Control blocks
  - ○ Procedure
  - ○ Data analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104671.

ACKNOWLEDGMENTS

This work was supported by the National Institute on Deafness and Other Communications Disorders Grant R01 DC-013543 (A.J.S.).

AUTHOR CONTRIBUTIONS

Conceptualization, A.J.S. and K.C.B.; Methodology, A.J.S, K.C.B., H.B., and M.G.G.; Resources, A.J.S., K.C.B., and L.M.M.; Investigation, M.G.G.; Formal Analysis, K.C.B., M.G.G., and Y.Y.; Writing – Original Draft, A.J.S. and K.C.B. Writing – Review & Editing, A.J.S., K.C.B., H.B., L.M.M., M.G.G., and Y.Y.; Visualization, A.J.S., K.C.B., H.B., M.G.G., L.M.M., and Y.Y.; Supervision, A.J.S. and K.C.B.; Funding Acquisition, A.J.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 28, 2022
Revised: May 6, 2022
Accepted: June 21, 2022
Published: July 15, 2022

REFERENCES

Abbott, N.T., and Shahin, A.J. (2018). Cross-modal phonetic encoding facilitates the McGurk illusion and phonemic restoration. J. Neurophysiol. *120*, 2988–3000. https://doi.org/10.1152/jn.00262.2018.

Besle, J., Fort, A., Delpuech, C., and Giard, M.H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. Eur. J. Neurosci. *20*, 2225–2234. https://doi.org/10.1111/j.1460-9568.2004.03670.x.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A.A. (2009). The natural statistics of audiovisual speech. PLoS Comput. Biol. *5*, e1000436. Epub 2009 Jul 17. https://doi.org/10.1371/journal.pcbi.1000436.

Chen, L., and Vroomen, J. (2013). Intersensory binding across space and time: a tutorial review. Atten. Percept. Psychophys. *75*, 790–811. https://doi.org/10.3758/s13414-013-0475-4.

De Kock, R., Gladhill, K.A., Ali, M.N., Joiner, W.M., and Wiener, M. (2021). How movements shape the perception of time. Trends Cognit. Sci. *25*, 950–963. https://doi.org/10.1016/j.tics.2021.08.002.

Elff, M. (2021). Mclogit: Multinomial Logit Models, With Or Without Random

Effects Or Overdispersion. R package version 0.8.7.3.

Kohl, M. (2020). MKinfer: Inferential Statistics. R package version 0.6. http://www.stamats.de.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. Nature *264*, 746–748. https://doi.org/10.1038/264746a0.

Recanzone, G.H. (2003). Auditory influences on visual temporal rate perception. J. Neurophysiol. *89*, 1078–1093. https://doi.org/10.1152/jn.00706.2002.

Schwartz, J.L., and Savariaux, C. (2014). No, there is no 150ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. PLoS Comput. Biol. *10*, e1003743. https://doi.org/10.1371/journal.pcbi.1003743.

Sekuler, R., Sekuler, A.B., and Lau, R. (1997). Sound alters visual motion perception. Nature *385*, 308. https://doi.org/10.1038/385308a0.

Shahin, A.J., Backer, K.C., Rosenblum, L.D., and Kerlin, J.R. (2018). Neural mechanisms underlying cross-modal phonetic encoding. J. Neurosci. *38*, 1835–1849. https://doi.org/10.1523/JNEUROSCI.1566-17.2017.

Shams, L., Kamitani, Y., and Shimojo, S. (2000). What you see is what you hear. Nature *408*, 788. https://doi.org/10.1038/35048669.

Shams, L., Kamitani, Y., Thompson, S., and Shimojo, S. (2001). Sound alters visual evoked potentials in humans. Neuroreport *12*, 3849–3852. https://doi.org/10.1097/00001756-200112040-00049.

Shams, L., Kamitani, Y., and Shimojo, S. (2002). Visual illusion induced by sound. Brain Res. Cognit. *14*, 147–152. https://doi.org/10.1016/s0926-6410(02)00069-1.

Stein, B.E., London, N., Wilkinson, L.K., and Price, D.D. (1996). Enhancement of perceived visual intensity by auditory stimuli: a psychophysical analysis. J. Cognit. Neurosci. *8*, 497–506. https://doi.org/10.1162/jocn.1996.8.6.497.

Sumby, W.H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. *26*, 212–215. https://doi.org/10.1121/1.1907309.

van Wassenhove, V., Grant, K.W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. Proc. Natl. Acad. Sci. USA *102*, 1181–1186. https://doi.org/10.1073/pnas.0408949102.

Vroomen, J., and de Gelder, B. (2000). Sound enhances visual perception: cross-modal effects of auditory organization on vision. J. Exp. Psychol. Hum. Percept. Perform. *26*, 1583–1590. https://doi.org/10.1037/0096-1523.26.5.1583.

Vroomen, J., and de Gelder, B. (2004). Temporal ventriloquism: sound modulates the flash-lag effect. J. Exp. Psychol. Hum. Percept. Perform. *30*, 513–518. https://doi.org/10.1037/0096-1523.30.3.513.

Welch, R.B., and Warren, D.H. (1980). Immediate perceptual response to intersensory discrepancy. Psychol. Bull. *88*, 638–667. https://doi.org/10.1037/0033-2909.88.3.638.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw Behavioral Responses	https://data.mendeley.com/datasets/sxrchpgvpg/1.	
Example Stimuli	https://data.mendeley.com/datasets/sxrchpgvpg/1.	
Software and algorithms		
In-house stimulus presentation code	https://www.neurobs.com/	
In-house Matlab parse code	https://www.mathworks.com/products/matlab.html	

RESOURCE AVAILABILITY

Lead contact

Questions and requests for information and data/code should be directed to the corresponding author (ashahin@ucmerced.edu).

Materials availability

Stimulus examples and original raw behavioral response data as text files can be found on https://data.mendeley.com/datasets/sxrchpgvpg/1. The site includes a text file explaining the variables in the text files.

Data and code availability

- Stimulus examples and behavioral response data can be found on https://data.mendeley.com/datasets/sxrchpgvpg/1 (https://doi.org/10.17632/sxrchpgvpg.1). The items are available to readers for download.
- Code used for stimulus presentation and code used to analyze the data can be requested from the lead contact.
- Any additional information needed to assess the current behavioral data can be obtained via contacting the lead contact.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Participants

Twenty-four individuals (>18 years of age, M = 21.25 years, SD = 4.33; 16 females) participated in the study. All participants reported native/fluent English background, normal hearing, right-handedness, and normal/corrected-to-normal vision. Only one participant was a non-native English speaker, who reported that they began learning English before the age 10 years. Prior to participation, participants provided written informed consent, and filled out a questionnaire about their language background, educational level, handedness and neurological history. Informed consent was obtained from the participants prior to participation, and Experimental methods were carried out in accordance with protocols approved by the Institutional Review Board (IRB) of the University of California, Merced. Monetary compensation was given to all participants for their participation.

METHOD DETAILS

Stimuli *experimental blocks*

The stimuli consisted of 3-s audiovisual video clips created by merging silent videos and corresponding audio clips of a female talker uttering trisyllabic words. The visual portions of the videos were cropped to only show the talker's face below the eyes to the bottom of the neck. There were three audiovisual stimulus conditions: *unchanged*, *paused*, *skipped*. Each condition consisted of 50 videos. To control for context influences, the words were different across the three conditions. The visual portions of the videos of all three conditions were intact. The three conditions only varied in the acoustic portions of the videos. In the *unchanged* condition, the audio portions of the words were intact, uttered in normal form. In the *paused* condition, the audio portions of the words contained two 100-ms silent segments inserted in the audio 200-ms apart, giving the perception that the audio is pausing. The onset of the first silence insertion began about a third of the way into the beginning of the word, followed by a 200-ms of the remaining speech signal, followed by another 100-ms silence, and finally followed by the rest of the speech signal. These parameters were based on the senior author's own judgment, as to what parameters might induce the strongest *pause* illusion. In the *skipped* condition, two speech segments were excised from the audios to induce a perception that the audio is skipping. A 100-ms segment was excised, beginning third of the way of the word, followed by a 133-ms of the speech, followed by another 100-ms excision, and finally followed by the rest of the word. Again, these arbitrary parameters were based on the senior author's judgment. Finally, we also included 5 *catch* trials in which the audio and corresponding visual portions of the videos were excised (*AV-skip*), and 5 *catch* trials in which the audio contained silent portions and the visual portions contained paused (repeated frames) corresponding to the same time windows of the audio's silent segments (*AV-pause*). The words used in the *catch* trials were different from the original three conditions. The reason for inclusion of the *catch* trials, was to monitor the participants' task vigilance. All audio clips were group-normalized in Adobe Audition to the same sound pressure level of −25 dB and presented at an average of 56 dB SPL (measured at head location). The audios can be accessed via the following link: https://data.mendeley.com/datasets/sxrchpgvpg/1.

Control blocks

The manipulations of the audiovisual stimuli in the control blocks were reversed across the two modalities for the three original conditions. The same words were used, as well as *catch* trial conditions.

Procedure

The participants sat about a meter from a 27-inch monitor and performed the task in a sound-attenuated booth. The task consisted of 5 blocks: Two practice blocks, 2 experimental blocks, and 2 control blocks. The practice blocks, one for the main and one for the control, consisted of 6 trials each: 2 *unchanged*, 2 *paused*, and 2 *skipped* trials. The main and control blocks each consisted of 75 randomly ordered trials for a total of 150 trials: 50 *unchanged*, 50 *paused*, and 50 *skipped* trials. The main and control block order was counterbalanced across subjects—12 participants received the main blocks first, and the other 12 participants received the control blocks first. Visual stimuli were presented through the 27-inch monitor, and auditory stimuli were presented through two loudspeakers located on the left and right sides (−/+45°) of the monitor. The sound was heard as coming directly from midline (0°). The participants were instructed to indicate whether they see the visual portion of the videos as *skipping or speeding-up*, *pausing* (*"freezing"*) or *"slowing-down"*, or neither *skipping* nor *pausing* (*unchanged*), by pressing the left arrow, right arrow, or down arrow, respectively, on the keyboard.

Data analysis

Logfiles of the participants' responses were parsed in MATLAB 2015 (MathWorks, Natick MA). Individual mean percept percentages were calculated as the number of responses for each percept *(e.g., "skip")* divided by all responses within each stimulus type (e.g., *skip*) x 100. Since there were three response options, chance level was therefore 33%.

QUANTIFICATION AND STATISTICAL ANALYSIS

We originally planned to run analysis of variance (ANOVA) to examine the percentage difference across modalities and conditions and run pairwise comparisons, but due to the violation of analysis of variance (ANOVA) assumptions of independence of the data, e.g., the percentage of *A-pause* as *"V-pause"* plus *A-pause* as *"V-skip"* plus *A-pause* as *"V-unchanged"* is equal to 1 (see Penn State University. Applied Statistics. 10.2.1 - ANOVA Assumptions. Retrieved 5/25/2022 From: https://online.stat.psu.edu/stat500/lesson/10/10.2/10.2.1#:~:text=There%20are%20three%20primary%20assumptions,The%20data%20are%20independent), we decided against an ANOVA. Rather, Separate bootstrap t-tests were performed to test whether the average percentage for each illusory percept of interests was significantly greater or lower than the average percentage for the same illusory percept in the *unchanged* condition, accounting for non-normality of the population of measurements (Kohl, 2020). Observations were considered dependent within individuals. Therefore, mixed-effects logistic regression models were computed on the single-trial data (i.e., observations) and used to compare the intact conditions or *catch* trials across the main and control blocks (i.e., fixed effects), while allowing a random intercept to account for the dependence of observations within individuals. The R package "MKinfer "(Kohl, 2020) and "mclogit" (Elff, 2021) were used to perform the bootstrap t-tests and mixed-effects models, respectively.