Journal of Experimental Psychology: General

Salient Features of Task-Irrelevant Continuous Speech Distort Subjective Time

Ashley E. Symons, Fred Dick, Lori Holt, and Adam Tierney

Online First Publication, May 18, 2026. https://dx.doi.org/10.1037/xge0001946

CITATION

Symons, A. E., Dick, F., Holt, L., & Tierney, A. (2026). Salient features of task-irrelevant continuous speech distort subjective time. *Journal of Experimental Psychology: General*. Advance online publication. https://dx.doi.org/10.1037/xge0001946

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2026 American Psychological Association
ISSN: 0096-3445

https://doi.org/10.1037/xge0001946

Salient Features of Task-Irrelevant Continuous Speech Distort Subjective Time

Ashley E. Symons[1], Fred Dick[2], Lori Holt[3], and Adam Tierney[4]

[1] Department of Psychology, Royal Holloway, University of London
[2] Department of Experimental Psychology, University College London
[3] Department of Psychology, The University of Texas at Austin
[4] Department of Psychological Sciences, Birkbeck College, University of London

Computational models of auditory salience predict that acoustic change and divergence from prediction increase the salience of sound streams. Confirming these predictions, prior research has shown that acoustic change and unpredictable sound features are linked to increases in physiological arousal and disruption of concurrent task performance. However, it remains unclear whether linguistic features, such as phonemic and lexical/semantic surprisal, help drive attentional orienting or whether instead attentional capture takes place prior to linguistic analysis. To address this question, we introduce a new technique for assessing attentional capture by naturalistic task-irrelevant speech. In this paradigm, participants tap to a metronome while ignoring a spoken passage from an audiobook. Salient features of the task-irrelevant speech capture attention, increase arousal, and expand subjective time, leading to shifts in tap timing. We show that distortions of subjective time are driven not only by acoustic change but also by phonemic surprisal. Thus, attentional orienting to sound takes place after the initial stages of linguistic analysis.

Public Significance Statement

We measured the speed of time perception while participants ignored distracting speech. We find that when listeners' predictions about upcoming speech sounds fail, the subjective passage of time slows down. This suggests that people make linguistic predictions even when ignoring speech and that prediction errors capture attention.

Keywords: attention, speech, time

Supplemental materials: https://doi.org/10.1037/xge0001946.supp

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Imagine that you are in a coffee shop, trying to work on a grant proposal. The ambient noise of silverware clinking and coffee being assembled recedes into the background as you focus your mind. Then, behind you, a conversation turns heated: A couple begins to argue, their voices suddenly louder, higher in pitch, and spiked with emotional words. Despite your best intentions, your attention drifts away from your proposal and you begin to eavesdrop. As this emotional conversation captures your attention, your physiological arousal increases: Your pupils dilate, your skin sweats, your pulse quickens, and your perception of time expands. This is a common experience because speech is particularly good at capturing our attention. However, it remains an open question which factors cause speech to capture attention and which cause speech to fade into the background.

Researchers have developed several computational models of the factors that drive attentional capture by sound. However, these models have been developed to apply to sound in general and so cannot capture speech-specific factors such as phonology or lexical

Sarah Brown-Schmidt served as action editor.
Ashley E. Symons https://orcid.org/0000-0001-5980-6752
Fred Dick https://orcid.org/0000-0002-2933-3912
Lori Holt https://orcid.org/0000-0002-8732-4977
Adam Tierney https://orcid.org/0000-0002-7624-6918
All processed data and analysis scripts are available on the Open Science Framework at https://osf.io/pc5tw/. Preliminary results from Experiment 1 were presented as a poster at the Society for the Neurobiology of Language Conference in 2023.

This research is funded by the National Institutes of Health (Grant NIH R01DC017734 awarded to Fred Dick). The authors thank the participants. The authors also thank Ed Lalor for providing the stimuli and stimulus information vectors used in Experiments 1, 2, and 4.

Ashley E. Symons played a lead role in data curation, formal analysis, investigation, methodology, project administration, software, and visualization and an equal role in conceptualization and writing–original draft. Fred Dick played a lead role in funding acquisition and a supporting role in conceptualization, formal analysis, supervision, and writing–review and editing. Lori Holt played a supporting role in writing–review and editing and an equal role in funding acquisition. Adam Tierney played a lead role in supervision and writing–review and editing, a supporting role in formal analysis, funding acquisition, methodology, and software, and an equal role in conceptualization and writing–original draft.

Correspondence concerning this article should be addressed to Ashley E. Symons, Department of Psychology, Royal Holloway, University of London, Egham Hill, Egham, TW20 0EX, United Kingdom. Email: ashley.symons@rhul.ac.uk

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

semantics. Instead, these models have focused on modeling changes in salience over time in complex sound streams due to acoustic factors. For example, some bottom-up models have created salience maps with center-surround inhibition in time-frequency "space" (Duangudom & Anderson, 2007; Kalinli & Narayanan, 2007; Kayser et al., 2005), inspired by vision research using eye-tracking data as ground truth (Niebur et al., 2002). These models predict that sudden acoustic changes across multiple dimensions (amplitude, pitch, spectral shape) will be linked to transient increases in attentional capture. Other contextual models track dynamic changes in feature-specific deviance from prediction relative to local and longer term statistics (Kaya & Elhilali, 2014; Tsuchida & Cottrell, 2012). These models predict that moments of high unpredictability within a speech stream will be followed by short-term capture of attention.

Predictions of auditory salience models have been tested via behavior and physiology. One simple method of assessing salience used to validate computational models is to ask participants how salient a given sound or auditory scene is (Duangudom & Anderson, 2007; Kalinli & Narayanan, 2007) or to ask participants which of two competing scenes is more salient (Huang & Elhilali, 2017). This research has found that salience ratings are greater after change along several acoustic dimensions, including loudness, pitch, and spectral shape, as well as when a sound stream's acoustic characteristics diverge from the distributional statistics of the surrounding context. However, this approach relies upon participants having consistent and valid interpretations of the word/concept "salience." An alternate approach is to examine the effects of presentation of a sound stream on performance of a concurrent task, such as serial visual short-term memory (Jones & Macken, 1993). Visual memory is disrupted more by sound streams featuring acoustic change, and larger changes are linked to greater disruption (Jones et al., 2000; Schlittmeier et al., 2012). Moreover, unpredictable changes cause more disruption of performance than predictable changes (Bell et al., 2012, 2019). However, disruption of performance is not a pure measure of attentional capture, because it can potentially reflect interference with preconscious automatic processes (such as processing of serial order) as well as divergence of attention (Hughes, 2014).

An alternate approach to studying capture of attention by sound streams is to measure the physiological components of attentional orienting. Capture of attention by a particularly salient sound is accompanied by an increase in arousal that prepares the listener for action (Sokolov, 1963). These arousal effects can be a confounding factor when investigating attentional capture by assessing disruption of behavior, as distraction and arousal can have opposite effects on performance (Bonmassar et al., 2023; Masson & Bidet-Caulet, 2019). However, arousal can be assessed more directly by measuring physiological responses such as pupil dilation, skin conductance, and magnetoencephalography/electroencephalography. For example, sound intensity is linked to the degree of pupil dilation (Antikainen & Niemi, 1983; Liao et al., 2016) and the amplitude of the galvanic skin response (Barry, 1975). The degree of acoustic modulation is also linked to the extent of microsaccadic inhibition (Zhao, Yum, et al., 2019), pupil dilation (Bala & Takahashi, 2000; Marois et al., 2018), involuntary peripheral muscle responses (Schultz et al., 2021), decreased neural phase-locking to target stimuli (Huang & Elhilali, 2020), and the size of the P3a, an event-related potential component thought to reflect attentional orienting (Berti et al., 2004; Rinne et al., 2006). These physiological responses are not only driven by acoustic change but also factor in the surrounding context: Unpredictable stimuli lead to greater changes in pupil dilation (Friedman et al., 1973; Liao et al., 2018; Milne, Zhao, et al., 2021; Qiyuan et al., 1985; Southwell et al., 2017; Zhao, Chait, et al., 2019) and larger neural responses (Kaya et al., 2020).

In summary, computational modeling and behavioral/physiological research have demonstrated that acoustic change and unpredictability are linked to disruption of behavior and increased arousal. Acoustic factors alone, however, may not be sufficient to explain why certain sounds capture attention. Task-irrelevant comprehensible speech, for example, interferes with task performance more than acoustically matched nonspeech sounds (Dorsi et al., 2018; Le Compte et al., 1997; Little et al., 2010; Viswanathan et al., 2014), suggesting that certain linguistic factors additionally play a role in driving attentional orienting. One possible explanation for why speech can so effectively capture attention is that it contains probabilistic regularities on many different levels, including phonemic and semantic patterns, leading to predictions that capture attention when not fulfilled. However, modeling has not addressed the question of whether unpredictability of linguistic features can lead to attentional orienting. This question has also largely not been addressed experimentally, either using physiological or behavioral measures. An important exception is Röer et al. (2019), who found that semantically unexpected words can interfere with visual short-term recall. However, as mentioned above, interference of an auditory stimulus with visual recall can reflect either attentional capture or interference-by-process. For example, as suggested by Röer et al., the sequence processing necessary for chunking during visual recall could overlap cognitively with the process of integrating a word with its preceding semantic context. Moreover, Röer and Cowan (2021) found that unexpected words in a distractor stream do not interfere with comprehension of a target speech stream. It remains, therefore, an open question whether linguistic surprisal in a task-irrelevant stream of speech can lead to attentional orienting.

Here, we demonstrate a method of tracking attentional capture by task-irrelevant speech that can be used to assess the salience of phonemic and semantic surprisal while ruling out the influence of interference-by-process. This approach takes advantage of a well-documented link between increased arousal and expansion of subjective time. Expanded subjective time has been demonstrated due to a wide variety of experimental manipulations of arousal, including administration of methamphetamine to rats (Maricq et al., 1981), emotional content of stimuli (Droit-Volet & Meck, 2007; Droit-Volet et al., 2004; Gil & Droit-Volet, 2012; Lake et al., 2016), breath-holding (Schwarz et al., 2013), artificially raised body temperature (Wearden & Penton-Voak, 1995), and presentation of simple fluctuating stimuli such as clicks and flashes (Buffardi, 1971; Droit-Volet & Wearden, 2002; Ortega & López, 2008; Penton-Voak et al., 1996; Wearden et al., 1999). Moreover, the rate of subjective passage of time and pupil size have been shown to correlate in monkeys (Suzuki et al., 2016). These findings are compatible with models of subsecond time perception featuring an internal central clock (or clocks), which can vary in speed due to changes in internal state (Allman & Meck, 2012; Allman et al., 2014; Gibbon et al., 1984; Merchant et al., 2013). Assessing subjective time, therefore, enables measurement of arousal-induced task bias separately from task performance, which reflects a complex combination of attention, arousal, and process-based interference.

Prior research on arousal and bias in internal timing has presented single short sound events and assessed retrospective time perception. However, we have developed a technique that enables the assessment of ongoing subjective time throughout presentation of a complex sound stream. Participants are asked to tap to the beat of a 2-Hz click track while ignoring the presentation of distracting sounds. An auditory rather than visual pacing signal is used—that is, clicks rather than flashes—because participants have been reported to tap more consistently to auditory stimuli (Chen et al., 2002), and so this choice minimizes noise in our data due to intrinsic synchronization variability. Synchronized tapping requires participants to keep track of time so that an upcoming movement can be planned to align with the next click. In a previous article (Symons et al., 2024), we showed that presentation of distracting sounds and sound changes led to an expansion of subjective time, causing participants to wait for less time before making their next movement. This finding is conceptually similar to the filled duration illusion, in which silent intervals are perceived as being shorter in duration than intervals filled with sensory events (Buffardi, 1971; Ortega & López, 2008; Wearden et al., 2007). A likely explanation is that unexpected sounds and sound changes lead to increased arousal, speeding up internal pacemakers and expanding subjective time (Gibbon et al., 1984). Importantly, larger acoustic changes led to greater temporal distortions: A one-semitone pitch change, for example, did not affect tap timing, but a six-semitone pitch change did, suggesting that the shift in timing was driven by sound salience rather than simple perception of acoustic change. These findings were consistent across online and in-lab participant samples, suggesting that this online paradigm can be used to accurately measure subtle tapping shifts.

This previous study used relatively simple sounds (e.g., complex tones and white noise bursts) that allowed for precise control over variations in the acoustic features of interest. The use of synthesized sounds enabled us to vary individual acoustic dimensions while keeping the stimuli otherwise constant. However, it remains an open question to what extent those results generalize to more naturalistic listening scenarios where sounds vary across multiple acoustic and linguistic features simultaneously. To address this question, here, we used this synchronized tapping paradigm to investigate capture of attention by task-irrelevant naturalistic speech. Participants were asked to tap to the beat of a 2-Hz click track while ignoring a continuous stream of narrative speech (an audiobook recording of "The Old Man and the Sea"; Broderick et al., 2018; Di Liberto et al., 2015; Teoh et al., 2019). The degree to which listeners' tapping deviated from the beat of the click track (tapping asynchrony) provided a measure of temporal distortion. Based on prior work (Symons et al., 2024), we predicted that salient acoustic changes, including changes in intensity, pitch, and spectral shape, would increase autonomic arousal, leading to an overestimation of the passage of time and more negative asynchronies (earlier tapping). To test whether temporal distortions could be elicited by linguistic unpredictability, we computed measures of word frequency, phoneme surprisal, and semantic surprisal, features that have been shown to elicit changes in neural tracking of continuous speech (Broderick et al., 2018, 2022; Gillis et al., 2021; Weissbart et al., 2020).

Experiment 1

Method

Participants

A sample of 101 participants between the ages of 20 and 41 (M = 28.34 years, SD = 5.83; 65 female, 36 male, zero nonbinary) was recruited from the Prolific (https://www.prolific.com/) online recruitment platform. Due to the National Institutes of Health (NIH) funding requirements, data on race and ethnicity were collected. We placed no geographic restrictions on Prolific, and therefore, racial and ethnic categories may not have applied to participants outside of the United States. However, we report them here for completeness: From the original sample, 99 participants reported their ethnicity as not Hispanic or Latino and two preferred not to report. Forty-five participants were Black or African American, 43 were White, five were Asian, seven were more than one race, and one participant preferred not to report.

Automated screening procedures were set to ensure that participants spoke English as their native language and had no known hearing impairments. The experiment was conducted using the online experiment platform Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Participants were required to complete the experiment on a desktop or laptop with Google Chrome as the web browser and instructed to wear headphones for the duration of the experiment. All experimental procedures were approved by the Ethics Committee in the Department of Psychological Sciences at Birkbeck, University of London. Each participant provided informed consent and received monetary compensation for their participation at a standard rate.

Data from participants who did not report English as one of their native languages on a subsequent questionnaire were excluded from analysis (N = 15). To ensure that participants were complying with task instructions to tap along to the click track, we imposed an additional set of criteria for exclusion: did not tap at all during one or more run (N = 2), failed to synchronize with the clicks (N = 9) meaning they showed no significant clustering of phases across taps according to circ_rtest in the Matlab Circular Statistics Toolbox (Berens, 2009), or whose tapping variability (standard deviation of intervals between tap and click) was greater than 100 ms (N = 17). We also removed any participant whose responses were coarsely quantized (>15 ms quantization) due to the use of Bluetooth keyboards (which participants were explicitly requested not to use). Compared to wired keyboards, Bluetooth keyboards bin responses in much longer intervals and do not permit the temporal precision needed to measure small tapping shifts (~4–5 ms; Symons et al., 2024). To identify participants showing coarsely quantized responses, we binned the intertap intervals with an 0.1-Hz resolution, computed the autocorrelation function (0–100 ms lags), and then identified peaks in the autocorrelation function (minimum prominence = 0.3). This resulted in the removal of one additional participant. Last, we removed one participant who had <70% of valid taps for one or more excerpt. Valid taps were those occurring within 250 ms of a click (so if participants stopped tapping temporarily, these missed taps would be considered invalid) and falling within 3 standard deviations of their mean. The final sample consisted of 55 participants ages 20–40 (M = 29.09, SD = 6.17; 35 female, 20 male; racial/ethnic status using NIH reporting groupings: 55 not Hispanic or Latino, 20 Black or African American, 29 White, six more than one racial grouping). Of this

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

sample, 26 participants reported receiving musical training (ranging from 1 to 20 years). However, only 10 participants could be classed as musicians based on the 6-year criterion suggested by prior research (Zhang et al., 2020). Therefore, musical training was not factored into the analysis.[1] Additionally, 24 participants reported experience with other languages (with the age of second language acquisition ranging from 1 to 35 years).

Stimuli

Tapping Sequences. The continuous speech consisted of an audiobook recording of "The Old Man and the Sea" spoken by a professional male narrator with an American English accent (see Broderick et al., 2018; Teoh et al., 2019). The audiobook was divided into four excerpts, each 2–3 min in duration. Each excerpt was presented simultaneously with a 2-Hz isochronous click sequence (Figure 1). Clicks were broadband impulses spanning 10 time points (0.23 ms in duration with a 44.1-kHz sample rate). To ensure that the clicks were audible against the continuous speech, the peak amplitude of the speech was set to be 70% of the click amplitude. Prior to the onset of the continuous speech, 10 clicks presented against silence were added to allow participants time to synchronize their tapping with the click track.

Acoustic Features. The amplitude envelope and relative pitch of the speech recordings were obtained from previous research using this audiobook (Teoh et al., 2019). The amplitude envelope was extracted by filtering the speech waveform between 80 and 2,800 Hz and computing the absolute value of the Hilbert transform. The envelope was then low-pass filtered (cutoff = 30 Hz) and downsampled to 128 Hz. This provided a measure of amplitude level. In addition, we computed the change in amplitude across successive time points by calculating the derivative over time. Relative pitch was computed by extracting the fundamental frequency (F0) of the speech signal at 128 Hz and applying a z-transform to normalize the pitch based on the speaker's vocal range. In addition, we computed the change in relative pitch across successive time points by calculating the derivative over time. Spectral centroid and spectral flux were extracted from each speech recording using the *spectralCentroid* and *spectralFlux* functions in the Audio Toolbox implemented in Matlab (Version 2021a). Spectral centroid describes the center of gravity in the spectrum, while spectral flux measures the change in the spectrum over time. Both spectral measures were computed with a 23.4-ms window and 15.6-ms overlap between successive windows, resampled to a 128-Hz sampling rate, and z-scored.

Linguistic Features. All linguistic features were based off written transcripts of the audiobook. Phoneme surprisal was computed using a similar procedure to that reported in Brodbeck et al. (2018) First, phoneme onsets were automatically marked using the Gentle forced aligner (https://lowerquality.com/gentle/). Missing or incorrectly marked phonemes were adjusted by hand. Next, a phonetic dictionary with linked lexical frequencies was assembled by combining information from the SUBTLEX subtitle database (Brysbaert & New, 2009) and the Carnegie Mellon University Pronouncing Dictionary (https://www.speech.cs.cmu.edu/cgi-bin/cmudict). Any words in the text that were not present in SUBTLEX were given a frequency equal to the lowest possible value. For each phoneme, we calculated two "cohort frequency values": First, we calculated the summed frequency of all words in the dictionary matching the set of phonemes spanning the beginning of the word to the current phoneme. Second, we calculated the sum of the frequency of all words matching the set of phonemes from the beginning of the word to the previous phoneme. Phoneme surprisal was calculated as the negative log2 of the ratio between the first and the second value. For the first phoneme of a word, phoneme surprisal was the negative log2 of the ratio between the summed frequency of all words in the dictionary beginning with that phoneme and the summed frequency of all words in the dictionary. Finally, phoneme surprisal was stored as a vector with values equal to the surprisal of each phoneme across the duration of the phoneme and zeros elsewhere. Nonzero portions of this vector were z-scored such that the mean of the nonzero values was 0 and the standard deviation was 1, to make sure that phoneme surprisal analysis did not simply reflect the difference between the presence versus absence of phonemes.

Word frequency was extracted from the SUBTLEX-US database (Brysbaert & New, 2009). Word onsets were marked using Prosodylab-Aligner (Gorman et al., 2011). These markings were obtained from previous studies using this audiobook (Broderick et al., 2018). For each speech recording, we downsampled the recording to 128 Hz and extracted the time points corresponding to each word onset and offset. A custom Matlab script then searched for each word in the database. Word frequency was stored as a vector with values equal to the frequency of each word, with the value only changing with the onset of the next word. Nonzero portions of the vector were z-scored prior to analysis. Words that were not found in the database were set to the minimum word frequency value in the database. Contractions such as "aren't" are included in the database without the apostrophe (e.g., "aren't" is listed as "arent"). However, it was not clear how contractions that form words when taking out the apostrophe (e.g., "I'll" or "we're") are represented in the database. Since these instances were rare in the speech recordings we used, these words were ignored in the analysis. A full list of words not found in the database or excluded from analysis here can be found in Supplemental Table S2.1.

The semantic surprisal measure was obtained from previous research using this audiobook (Anderson et al., 2024). The text corresponding to each speech recording was passed to OpenAI's GPT-2, which computed a single surprisal value for each word based on the preceding context (up to 1,024 words). Each value represented the negative log probability estimate of each word. Semantic surprisal values were time-aligned with word onset and stored as a vector with surprisal values lasting the duration of the word, with the value changing at the onset of the next word. Nonzero portions of the vector were z-scored prior to analysis.

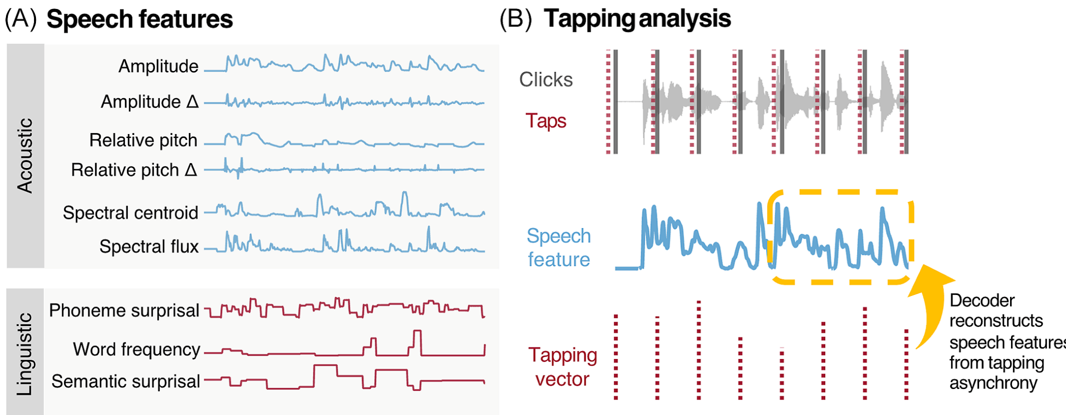
Procedure

Upon signing up to the study, participants were provided with a link to the experiment. After providing informed consent, participants completed a demographics questionnaire in which they reported their age, gender, language background, and musical experience. On-screen instructions were then provided. Participants were told that they would hear a series of clicks against some background sounds and instructed to tap to the beat of the clicks by pressing the "j" key on

[1] Given that this experiment was not designed to test effects of musicianship, we did not factor years of musical training into our primary analyses. However, because musical training affects beat synchronization (e.g., Thompson et al., 2015) and time perception (e.g., Mittal et al., 2024), we have included a preliminary analysis of musical training in the Supplemental Materials that can inform future work.

Figure 1

Display of Acoustic and Linguistic Features of Speech Stimulus and Diagram of Tapping Analysis



Note. On the left, examples of speech feature vectors are shown, which were extracted across the full duration of each excerpt. Acoustic and linguistic features were extracted from the audiobook and represented as vectors that were time-aligned with the tapping responses. On the right, a schematic of the tapping analysis is shown. Listeners tapped to the beat of a click track (dark gray lines) while ignoring continuous speech in the background (light gray). Taps are represented in dotted red vertical bars. Using the multivariate temporal response function toolbox, a decoder was trained to predict variations in speech features (amplitude envelope shown here) based on tapping asynchrony, which was represented as a single vector with values at each click time representing the difference between tap and click time. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

the keyboard while ignoring the background speech. An example sequence of clicks presented against silence was provided to allow participants to practice tapping to the clicks. During the main task, participants heard each excerpt with the order of the four runs (and thus the order of stimulus presentation) randomized across participants. At the end of the experiment, participants were asked whether they experienced any technical issues that could have affected their performance on the task. No technical issues were reported. This experiment lasted approximately 20 min.

Data Processing and Analysis

Tapping Asynchrony. Sound timing and participant response times were recorded in Gorilla (https://gorilla.sc/). Custom Matlab scripts extracted the sound offset (relative to the start of the run) and subtracted this from the known sound duration to measure the sound delay for each run (with each run consisting of one 2- to 3-min speech recording with clicks) and each individual. Participants' taps were extracted for each run and the difference between participants' tap time and the nearest click onset (tap–click asynchrony) was computed. The true asynchrony between participants' tap time and the click onsets could not be reliably recorded due to variations in the computer setup. To account for this variability, we subtracted the tap–click asynchrony at each time point from the mean tap–click asynchrony across the entire run. Instances in which there was no tap within ±250 ms of a click onset were classified as missing taps and excluded from analysis. Likewise, taps greater than 3 standard deviations from the participant's mean tapping asynchrony for a given run were removed from analysis. Of the participants included, the percentage of taps removed on this basis ranged from 0.28% to 5.41% ($M = 1.99\%$, $SD = 1.42\%$). Only participants with >70% of valid taps were included in the analysis. Taps within the first 5 s (10 clicks prior to the onset of the speech) were not included in the analysis.

Stimulus Reconstruction. To determine the relationship between the features of continuous speech and tapping asynchrony, a linear model was trained to reconstruct an estimate of each feature separately based on tapping asynchrony using the multivariate temporal response function (TRF) toolbox (Crosse et al., 2016) implemented in Matlab (Version 2023b). To do this, we treated the tapping data as a vector consisting of a series of impulses at the click time points and zeros at all other time points, with the amplitude of each impulse equal to the corresponding tapping asynchrony. Nonzero portions of the tapping vector were z-scored prior to analysis. This tapping vector was then used to predict the speech features in the 2 s preceding the click using the "backwards" option in the multivariate TRF toolbox. A doubly nested cross-validation procedure was used to identify the optimal regularization parameter and then test the model. First, we divided the data from the four runs into a training set consisting of three runs and a test set consisting of one run. Then a leave-one-out cross-validation procedure was conducted on the training data over time lags from zero to a maximum of 2 s, with the time lags selected based on previous research (Symons et al., 2024). This procedure was used to obtain the optimal regularization parameter within the range of 2^1 to 2^{10}. The model was then trained using the optimal regularization parameter. To evaluate the performance of the model, we determined the accuracy with which the model predicted speech features based on tapping asynchrony in the test data. The similarity between the predicted and observed data was computed using Spearman's correlation coefficient. This process was repeated four times, with each run taking its turn serving as the test data and the remaining three runs serving as the training data (e.g., Model 1 was trained on Runs 2–4 and tested on Run 1). The prediction accuracy (rho value) of the model and model coefficients were averaged across all four models.

Statistical Analysis. To test whether variations in tapping asynchrony predicted the stimulus features of interest, we conducted

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

a Monte Carlo analysis. During each of 1,000 iterations, we randomly shuffled the tapping data within each excerpt for each participant. We then ran the TRF analysis reported above on the shuffled tapping data, computing the resulting prediction accuracy for each participant. Next, we took the median prediction accuracy across participants, resulting in a null distribution of median cross-participant prediction accuracy over the 1,000 iterations. Finally, we compared the median of the prediction accuracy across participants for the original data to this null distribution. The *p* values, therefore, represent the probability of the observed rho value given the distribution of rho values obtained when shuffling the tapping data.

Prior work has shown that temporal distortions occur between 250 and 750 ms following salient acoustic changes (Symons et al., 2024). To examine the time course of temporal distortions elicited by variations in continuous speech, we examined the behavioral TRF, which represents the degree to which tapping asynchrony predicted variations in each feature at each time point (within the 2-s time window) preceding the tap. One-sample rank sum tests compared model coefficients to zero across participants at each time point within the 2-s time window preceding the tap to establish the time course of the tapping shift. The *p* values were corrected for multiple comparisons across time (257 time points; Benjamini & Hochberg, 1995).

There was a significant correlation between amplitude envelope and many of the other stimulus features (see Supplemental Figure S1.1). Therefore, any effects observed for the other stimulus features could be partially driven by amplitude. To ensure that the effects observed for other features were not solely driven by amplitude, we ran a follow-up analysis covarying out amplitude. To do this, we constructed a linear regression using amplitude to predict each of the other features and extracted the residuals from the model. We then conducted the same statistical analyses as above, but with the residuals from the regression model as the dependent variable.

Transparency and Openness

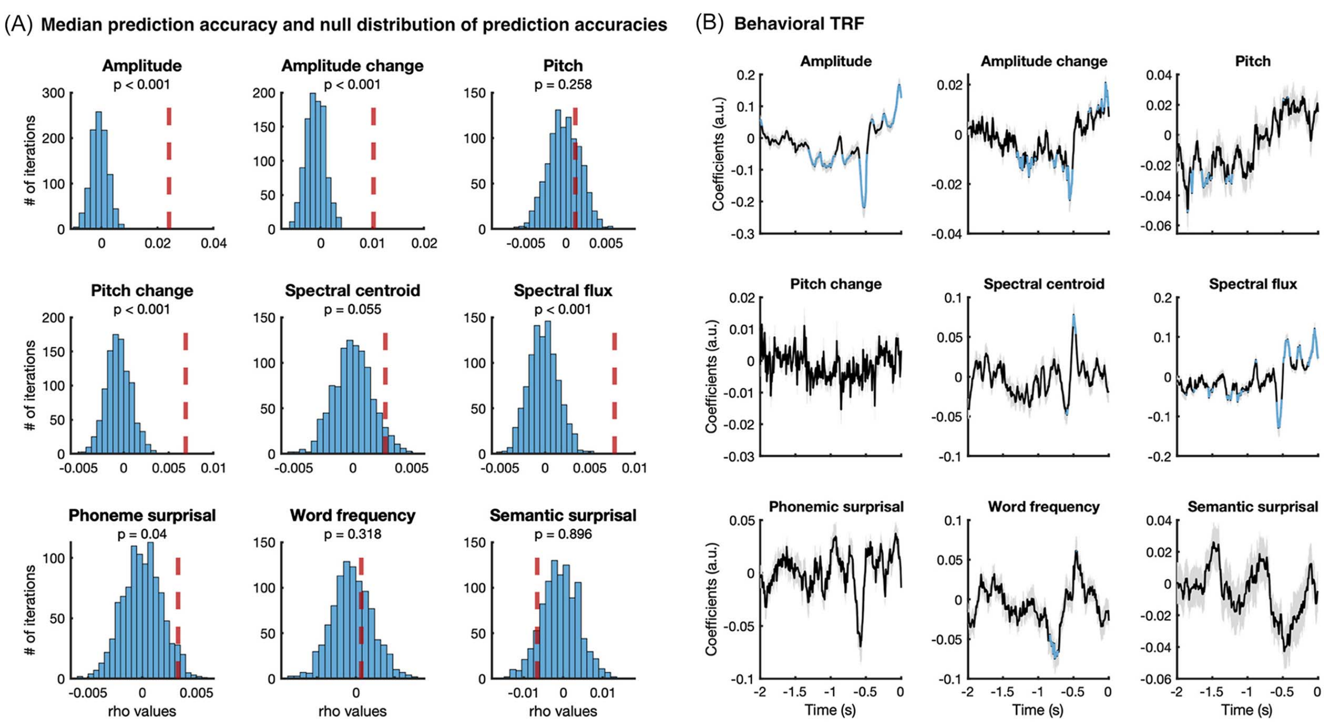
This study was not preregistered. The sample size was determined based on previous research (Symons et al., 2024). All data inclusion criteria, manipulations, and measures are reported here. Stimulus vectors, anonymous data, and analysis code are available on the Open Science Framework (https://osf.io/pc5tw/; Symons et al., 2026).

Results

Tapping asynchrony significantly predicted amplitude, amplitude change, pitch change, spectral flux, and phoneme surprisal (Figure 2A).

Figure 2

Experiment 1: Relationship Between Tapping Asynchrony and Features of Task-Irrelevant Speech



Note. Panel A shows the prediction accuracy. The dashed line shows the correlation coefficient (Spearman's rho) representing the relationship between the time series of each speech feature and the predicted time series based on tapping asynchrony as estimated by the multivariate TRF model. Histograms show the permutation-generated null distribution of rho values, representing the relationship between the time series of each speech feature and the time series predicted by the shuffled tapping data. The *p* values represent the probability of the observed rho value given the distribution of rho values obtained when shuffling the tapping data. Panel B shows the behavioral TRF. Coefficients (in a.u.) represent the degree to which tapping asynchrony predicts the speech feature at each time lag. Along the *x*-axis, the zero time lag indicates the onset of the click to which participants were attempting to synchronize. Along the *y*-axis, a positive coefficient indicates that a higher value (e.g., larger amplitude) is associated with later tapping while a negative coefficient indicates that a higher value is associated with earlier tapping. Thick blue lines represent lags at which the coefficients significantly differ from zero (with false discovery rate correction for multiple comparisons). TRF = temporal response function; a.u. = arbitrary units. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Across features, tapping asynchrony was most consistently linked to speech features occurring 550–600 ms before the click (Figure 2B). Moments in the speech featuring high amplitude, for example, were linked to earlier tapping roughly half a second later. Earlier tapping was also linked to moments of high acoustic *change* in the speech half a second earlier, including changes in amplitude, pitch, and spectral shape (frequency content). These results suggest that acoustic change increases physiological arousal, with a lag of around 500 ms, causing participants to experience an expansion of subjective time and, therefore, tap earlier.

Importantly, effects of speech characteristics on tapping asynchrony were not limited to acoustic measures but extended to linguistic characteristics: Low phonemic predictability also led participants to tap sooner. Therefore, failure of linguistic prediction led to faster tapping roughly a half second later, suggesting an expansion of subjective time due to increased arousal.

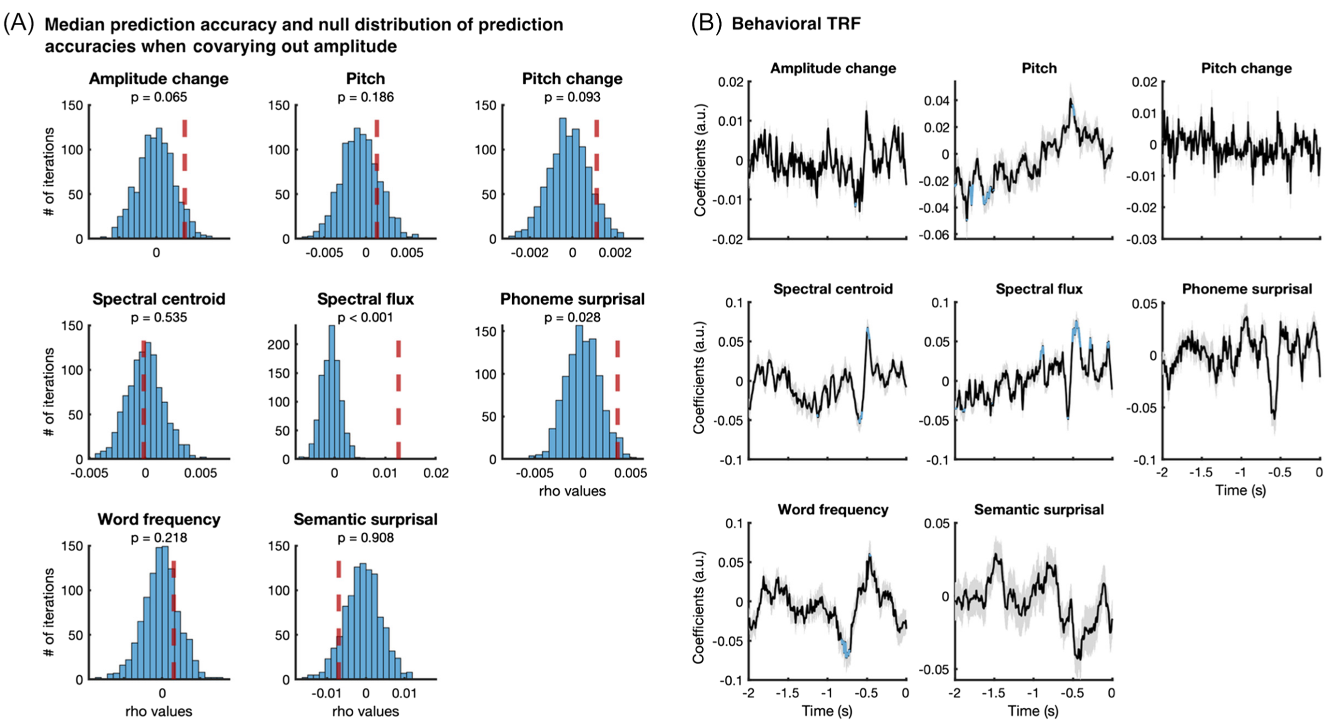
To ensure that the links between speech features and tapping speed were not simply driven by variations in amplitude, we ran a follow-up analysis covarying out amplitude. Figure 3 shows the median prediction accuracy versus a histogram of the null distribution of prediction accuracies, as well as model coefficients, when covarying for amplitude. Effects of amplitude and pitch change (both highly correlated with amplitude; see the Supplemental Materials) were no longer significant. However, even when covarying out amplitude, tapping asynchrony was significantly linked to both preceding spectral flux and phoneme surprisal. Both acoustic change and linguistic surprisal, therefore, continued to predict tapping speed, even once the effects of amplitude were controlled for.

Discussion

We find that acoustic changes in task-irrelevant speech, including changes in amplitude, pitch change, and spectral shape, are linked to distortions in subjective time, as measured with a synchronized tapping paradigm. Our prediction, based on the results of our previous experiment (Symons et al., 2024), was that acoustic changes would be linked to earlier tapping 250–750 ms later. The functions relating tapping asynchrony to acoustic change are roughly consistent with these predictions: For amplitude change, pitch change, and spectral flux, greater change was linked to earlier tapping between 500 and 1,250 ms later. This pattern suggests that acoustic change led to an increase in arousal arising within around 500 ms, expanding subjective

Figure 3

Experiment 1: Relationship Between Tapping Asynchrony and Features of Task-Irrelevant Speech (Covarying for Amplitude)



Note. Panel A shows the prediction accuracy when covarying out amplitude. Amplitude was covaried out by regressing each speech feature against amplitude and extracting the residuals. The dashed line shows the correlation coefficient (Spearman's rho) representing the relationship between the time series of each speech feature (with amplitude removed) and the predicted time series based on tapping asynchrony as estimated by the multivariate TRF model. Histograms show the permutation-generated null distribution of rho values, representing the relationship between the time series of each speech feature (with amplitude removed) and the time series predicted by the shuffled tapping data. The *p* values represent the probability of the observed rho value given the distribution of rho values obtained when shuffling the tapping data. Panel B shows the behavioral TRF. Coefficients (in a.u.) represent the degree to which tapping asynchrony predicts the speech feature after removing the contribution of amplitude at each time lag. Along the *x*-axis, the zero time lag indicates the onset of the click to which participants were attempting to synchronize. Along the *y*-axis, a positive coefficient indicates that a higher residual value (e.g., higher pitch after accounting for amplitude) is associated with later tapping while a negative coefficient indicates that a higher residual value is associated with earlier tapping. Thick blue lines represent lags at which the coefficients significantly differ from zero (with false discovery rate correction for multiple comparisons). TRF = temporal response function; a.u. = arbitrary units. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

time for around 750 ms before returning to baseline. We also find that sounds with greater amplitude are linked to earlier tapping in the same time range, suggesting that louder sounds are more salient, leading to greater attentional orienting, increased arousal, and expanded subjective time.

Importantly, we find that the link between tapping asynchrony and characteristics of task-irrelevant speech is not limited to acoustic features. There was a robust relationship between phonemic surprisal and asynchrony, such that greater surprisal was linked to earlier tapping. The time course of this effect closely aligned with the time course of the effect of amplitude on tapping; however, phoneme surprisal and amplitude only weakly correlated ($r_s = 0.07$), and the phoneme surprisal effect remained significant even after covarying for amplitude. This finding suggests that phonemic surprisal captures attention, increasing arousal and expanding subjective time. We did not find any significant relationship between semantic surprisal and time perception; however, this null result could also reflect a lack of statistical power and so should be interpreted with caution. Our finding that semantic surprisal does not affect tapping performance conflicts somewhat with the finding of Röer et al. (2019) that semantic unpredictability in speech can interfere with the performance of a concurrent visual serial memory task, but as the authors of that article suggest, this could reflect interference by process between semantic integration and tracking of serial order. Our results also conflict somewhat with Kothinti and Elhilali (2023), who found that semantic surprisal in nonlinguistic auditory scenes was a predictor of perceptual salience, as measured via salience ratings.

Our primary framework for explaining our results is that they reflect fluctuations in arousal, which have been shown to be linked to expansions and contractions in subjective time (Droit-Volet & Meck, 2007; Maricq et al., 1981; Schwarz et al., 2013; Wearden & Penton-Voak, 1995). However, an alternate possible explanation is that acoustic events in the period just before a click are perceptually fused with the click onset, resulting in a hybrid percept with an earlier time of onset. This could cause participants to perceive that their tapping is later than the hybrid perceived click, leading them to make their next movement earlier in time. Similar effects of integration of auditory events on the phase of synchronized tapping are demonstrated in Repp (2004), with a fixed window for temporal integration of around 120 ms. This perceptual fusion account could explain why the functions relating amplitude change, spectral flux, and amplitude level all peak just before the previous click (at 500 ms). However, this explanation would have difficulty explaining why phoneme surprisal is linked to tapping asynchrony, given that correlations between phoneme surprisal and acoustic features were rather weak ($r_s = 0.10$ or lower; see Supplemental Figure S1.1). Nevertheless, to rule out this explanation, we ran a follow-up experiment in which the clicks and the task-irrelevant speech were presented in separate ears, preventing perceptual fusion. This additional experiment also enabled us to determine the replicability of the shape of the functions relating speech features to tapping over time.

Experiment 2

Overview

Our results from Experiment 1 showed consistent shifts in synchronized tapping across multiple acoustic and linguistic features. The time course of this effect aligned with our previous research using simpler sounds (250–750 ms; Symons et al., 2024). However, because the clicks and speech were presented in the same ear, we could not rule out the possibility that the observed tapping shifts were driven by perceptual fusion of the speech with clicks as opposed to increases in arousal. Therefore, we conducted a second experiment aimed at (a) ruling out the possibility that tapping shifts were driven by perceptual fusion and (b) determining the replicability and generalizability of the relationship between stimulus dynamics and tapping shifts. To this end, Experiment 2 aimed to replicate the results of Experiment 1 in a new sample of participants and audiobook recordings. Listeners tapped to the beat of a click track while ignoring excerpts from "The Old Man and the Sea" (different from the excerpts used in Experiment 1). To determine whether the effects observed in Experiment 1 were driven by perceptual fusion, Experiment 2 presented clicks and speech in opposite ears. Half of the participants heard clicks in the left ear and speech in the right, while the other half of the participants heard clicks in the right ear and speech in the left. If the previous results were driven by perceptual fusion, there should be no relationship between tapping asynchrony and the features present in continuous speech. However, if variations in continuous speech distort internal timekeeping by changing physiological arousal, tapping asynchrony should predict acoustic and linguistic features even when the two streams are present in opposite ears.

Method

Participants

A sample of 194 participants between the ages of 18 and 40 (M = 30.91, SD = 5.99; 70 female, 120 male, four nonbinary) was recruited from the Prolific (https://www.prolific.com/) online recruitment platform. Data on race and ethnicity were not recorded since this experiment was not conducted with NIH funding. Automated screening procedures were set to ensure that participants spoke English as their native language and had no known hearing impairments. The experiment was conducted using the online experiment platform Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Participants were required to complete the experiment on a desktop or laptop with Google Chrome as the web browser. All participants were instructed to wear headphones for the duration of the experiment and completed a headphone screening test (Milne, Bianco, et al., 2021) to ensure that they were wearing headphones. The headphone test was needed in Experiment 2 because, unlike in Experiment 1, it was essential that clicks and speech were presented in opposite ears. All experimental procedures were approved by the Ethics Committee in the Department of Psychological Sciences at Birkbeck, University of London. All participants provided informed consent and received monetary compensation for their participation at a standard rate.

Data from participants who did not report English as one of their native languages on a questionnaire were excluded from analysis ($N = 3$). Participants who experienced technical issues with sound loading ($N = 20$) were excluded. Since the use of headphones was essential for this experiment, participants who achieved less than 6/6 on the headphone screening test (Milne, Bianco, et al., 2021) were also removed ($N = 26$). Following the same procedure as Experiment 1, participants who did not tap at all during one or more runs ($N = 1$), showed tapping variability >100 ms ($N = 37$), failed to synchronize with the clicks ($N = 7$), had keyboard quantization >15 ms ($N = 4$), or had fewer than 70% of valid taps during one or more run ($N = 2$) were excluded. The final sample consisted of 94 participants ages 19–40 (M = 32.05, SD = 5.46; 33 female, 59 male, two nonbinary). Of this sample, 44 participants reported receiving musical training (ranging from 1 to 15 years), with only 10 participants reporting at least 6 years of training (Zhang et al., 2020). For this reason, musical training was not factored into the analysis. Twenty-eight participants reported experience with other languages (with the age of second language acquisition ranging from 1 to 35 years).

Stimuli

As in Experiment 1, tapping sequences consisted of four 2-Hz isochronous click sequences (2–3 min in duration). The properties of the clicks were identical to Experiment 1. Each sequence was presented simultaneously with continuous speech excerpts from the same audiobook as Experiment 1 ("The Old Man and the Sea") but consisted of four different excerpts from those used previously. The peak amplitude of the continuous speech was set to be 70% of the click amplitude. In this experiment, clicks and speech were presented in opposite ears, with the ear in which clicks/speech were presented counterbalanced across participants. Acoustic (amplitude, amplitude change, relative pitch, pitch change, spectral centroid, spectral flux) and linguistic (word frequency, phoneme surprisal, semantic surprisal) features were extracted from the continuous speech following the same procedure as Experiment 1.

Procedure

The procedure for Experiment 2 was identical to Experiment 1.

Data Processing and Analysis

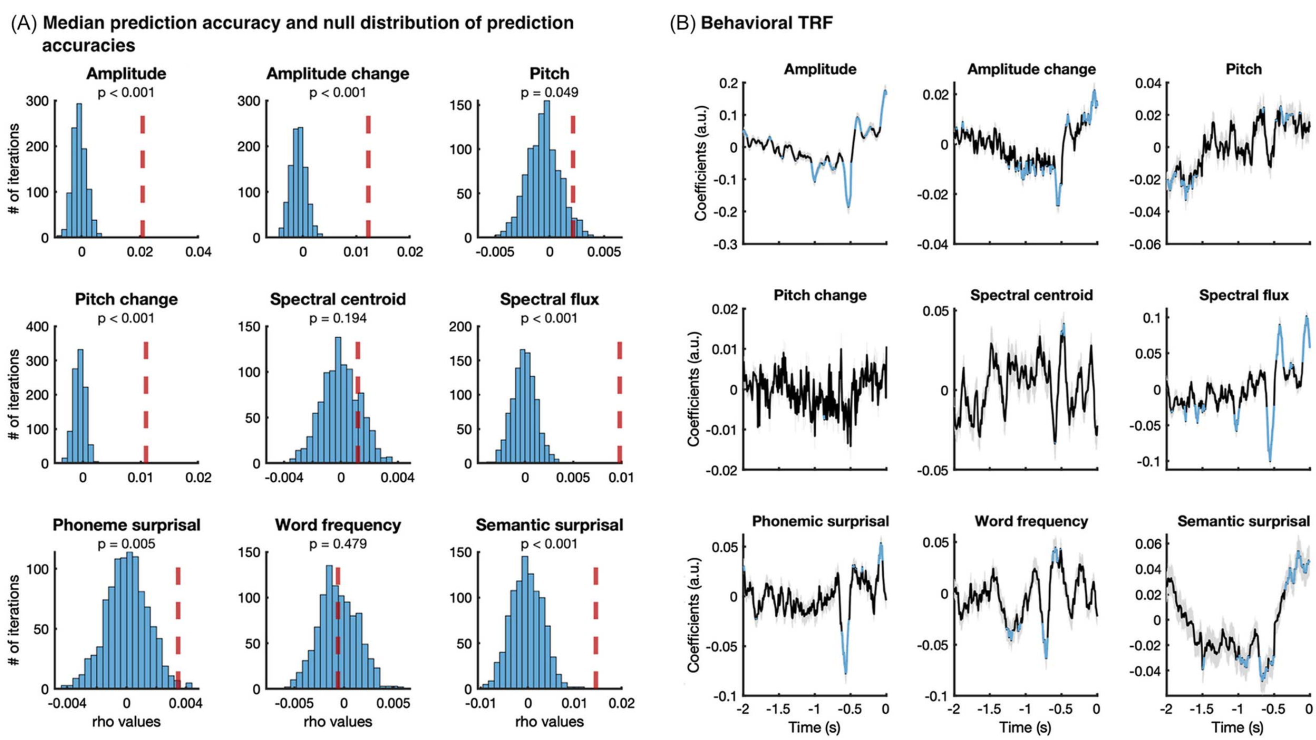
The data processing and analysis protocol was identical to Experiment 1. Of the participants included, the percentage of taps removed on our predefined exclusion criteria ranged from 0% to 8.08% (M = 1.92%, SD = 1.34%).

Results

When clicks and speech were presented in opposite ears, tapping asynchrony was most consistently linked to variations in the speech signal 550 ms before the click (Figure 4). As in Experiment 1, moments

Figure 4

Experiment 2: Relationship Between Tapping Asynchrony and Features of Task-Irrelevant Speech



Note. Panel A shows the prediction accuracy. The dashed line shows the correlation coefficient (Spearman's rho) representing the relationship between the time series of each speech feature and the predicted time series based on tapping asynchrony as estimated by the multivariate TRF model. Histograms show the permutation-generated null distribution of rho values, representing the relationship between the time series of each speech feature and the time series predicted by the shuffled tapping data. The p values represent the probability of the observed rho value given the distribution of rho values obtained when shuffling the tapping data. Panel B shows the behavioral TRF. Coefficients (in a.u.) represent the degree to which tapping asynchrony predicts the speech feature at each time lag. Along the x-axis, the zero time lag indicates the onset of the click to which participants were attempting to synchronize. Along the y-axis, a positive coefficient indicates that a higher value (e.g., larger amplitude) is associated with later tapping while a negative coefficient indicates that a higher value is associated with earlier tapping. Thick blue lines represent lags at which the coefficients significantly differ from zero (with false discovery rate correction for multiple comparisons). TRF = temporal response function; a.u. = arbitrary units. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

in the speech featuring high amplitude were linked to earlier tapping roughly half a second later. Earlier tapping was also preceded by time points in which the speech acoustics rapidly changed, including changes in amplitude, pitch, and spectral shape. As in Experiment 1, the relationship between speech features and tapping speed was not limited to acoustic factors. Linguistic surprisal, including both phonemic and semantic surprisal, led to earlier tapping, suggesting that failure of prediction during speech listening is linked to increased arousal and, therefore, expansion of subjective time.

Broadly, then, we replicated the Experiment 1 results, even though in this experiment the clicks and speech were in separate ears. There were some minor differences between the patterns of results in Experiment 2 versus Experiment 1: Contrary to Experiment 1, relative pitch and semantic surprisal were linked to tapping asynchrony in Experiment 2.

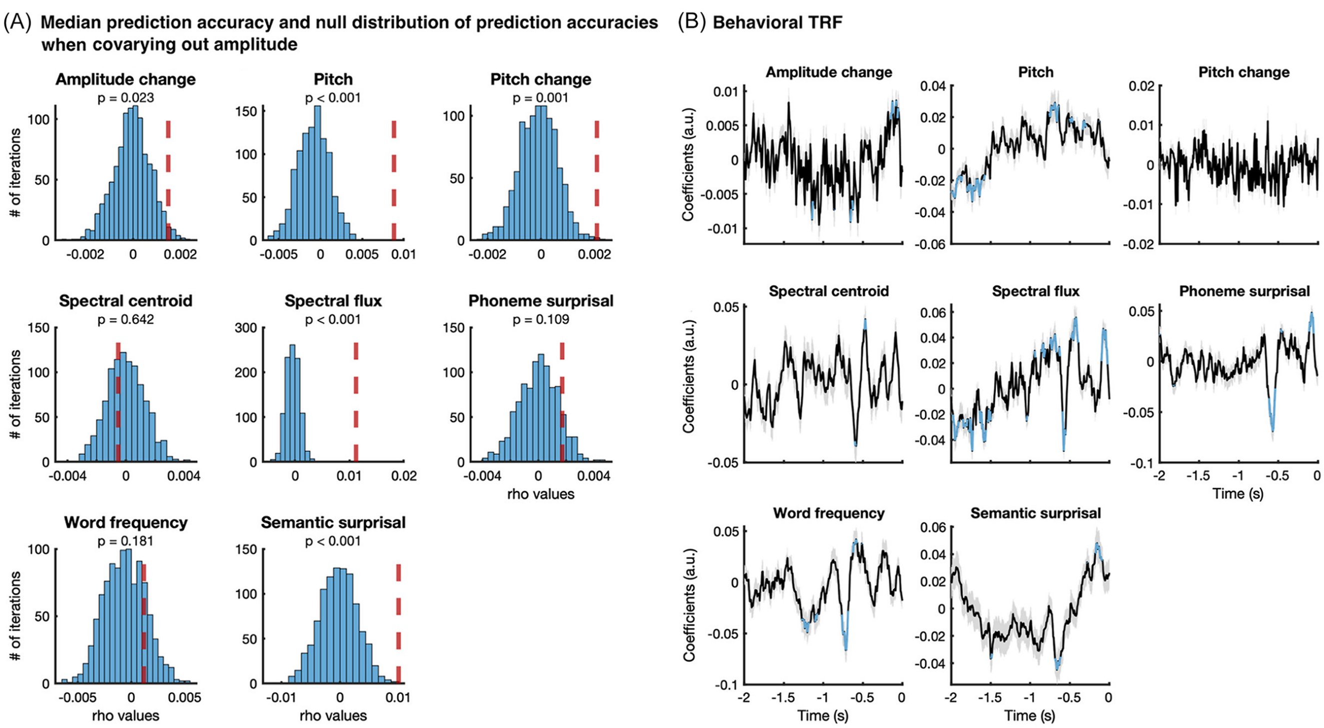
To ensure that the relationships between speech features and tapping asynchrony were not simply driven by variations in amplitude, we ran a follow-up analysis covarying out amplitude (Figure 5). Tapping asynchrony was significantly linked to preceding variations in amplitude change, relative pitch, pitch change, spectral flux, and semantic surprisal after accounting for variations in amplitude. Both acoustic change and linguistic surprisal, therefore, continued to predict tapping speed, even once the effects of amplitude were controlled for, though the effect of phoneme surprisal was no longer significant in this analysis.

Discussion

We find that when participants listen to speech in one ear while tapping to a click track in the other, moments in the speech that feature high amplitude, rapid acoustic change, and linguistic surprisal are followed by earlier tapping. That the link between speech features and tapping speed is present even when the target clicks and distracting speech are in opposite ears suggests that this relationship cannot be driven by perceptual fusion between the clicks and sound changes within a nearby temporal window (Repp, 2004). Nevertheless, in this experiment, we once again find that the function relating speech characteristics to tapping peaks at around 550 ms, just before the

Figure 5

Experiment 2: Relationship Between Tapping Asynchrony and Features of Task-Irrelevant Speech (Covarying for Amplitude)



Note. Panel A shows the prediction accuracy when covarying out amplitude. Amplitude was covaried out by regressing each speech feature against amplitude and extracting the residuals. The dashed line shows the correlation coefficient (Spearman's rho) representing the relationship between the time series of each speech feature (with amplitude removed) and the predicted time series based on tapping asynchrony as estimated by the multivariate TRF model. Histograms show the permutation-generated null distribution of rho values, representing the relationship between the time series of each speech feature (with amplitude removed) and the time series predicted by the shuffled tapping data. The p values represent the probability of the observed rho value given the distribution of rho values obtained when shuffling the tapping data. Panel B shows the behavioral TRF. Coefficients (in a.u.) represent the degree to which tapping asynchrony predicts the speech feature after removing the contribution of amplitude at each time lag. Along the x-axis, the zero time lag indicates the onset of the click to which participants were attempting to synchronize. Along the y-axis, a positive coefficient indicates that a higher residual value (e.g., higher pitch after accounting for amplitude) is associated with later tapping while a negative coefficient indicates that a higher value is associated with earlier tapping. Thick blue lines represent lags at which the coefficients significantly differ from zero (with false discovery rate correction for multiple comparisons). TRF = temporal response function; a.u. = arbitrary units. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

presentation of the previous click. Why are acoustic and linguistic factors particularly salient just before the presentation of a click?

In this paradigm, the task-relevant stimulus (i.e., the click) was perfectly predictable. As a result, participants knew that the time between clicks could only contain task-irrelevant sound. Participants may, therefore, have manipulated temporal attention to diminish the salience of any sounds presented in between clicks. One potential mechanism for this attentional weighting could be periodic modification of arousal. Shalev and Nobre (2022), for example, demonstrated that when temporally predictable stimuli were presented, tonic arousal was overall lowered, but increased briefly just before upcoming stimuli. In our experiment, then, participants may have lowered tonic arousal between clicks, tamping down the response to task-irrelevant speech. However, just before the onset of the next click, they may have increased arousal, making themselves vulnerable to attentional capture by speech features, including acoustic and linguistic change.

In Experiment 2, tapping shifts were linked to both phoneme surprisal and semantic surprisal. The effects of phoneme surprisal replicate effects observed in Experiment 1, though phoneme surprisal was not significant in the most conservative analysis in which clicks and speech were presented in opposite ears and amplitude was covaried out. By contrast, in Experiment 2, we found that semantic surprisal was significantly linked to subsequent tapping speed. Effects of semantic surprisal were not observed in Experiment 1 and should therefore be interpreted with caution. Given that the link between phonemic surprisal and tapping speed was weaker in Experiment 2 and was not significant when covarying for amplitude, we ran a third experiment to determine whether this relationship replicated in a different talker reading a different text.

Experiment 3

Overview

Results from Experiment 2 largely replicated those from Experiment 1, suggesting that acoustic and linguistic features of continuous speech can increase physiological arousal, resulting in a distortion of internal timekeeping. However, there were subtle differences across experiments and analyses as to which features were predicted by tapping shifts. Experiment 3 aimed to identify (a) which features were most consistently linked to shifts in tapping asynchrony and (b) whether the effects observed in Experiments 1 and 2 generalized to a novel speaker and a different narrative. In this experiment, a new sample of participants heard an audiobook recording of “Northern Lights” (Pullman, 1995) spoken by a female narrator. Based on the results of Experiments 1 and 2, we predicted that shifts in tapping asynchrony would predict variation in amplitude, amplitude change, pitch change, spectral flux, and phoneme surprisal.

Method

Participants

A sample of 98 participants between ages 20 and 66 (M = 38.52, SD = 9.69; 56 female, 42 male, zero nonbinary) was recruited from the Prolific (https://www.prolific.com/) and SONA recruitment platforms. No data on race or ethnicity were collected because the study was not conducted under NIH funding. All participants completed the study via Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Experimental procedures were approved by the Ethics Committee in the Department of Psychological Sciences at Birkbeck, University of London, and participants were provided with monetary compensation or course credit for their participation at a standard rate.

Data from participants who did not report English as their native language (N = 8) were excluded from analysis. Additionally, participants who tapped out of phase with the clicks (N = 5), had tapping variability >100 ms (N = 14), showed substantial (>15 ms) keyboard quantization (N = 4), or had fewer than 70% of valid taps (N = 2) were removed. (No participants failed to tap during a run.)

The final sample consisted of 65 participants (M_{age} = 39.92 years, SD = 10.13; 37 female, 28 male). Of this sample, 25 participants reported receiving musical training (ranging from 1 to 15 years). Only six could be classified as musicians based on a 6-year criterion (Zhang et al., 2020). Fifteen participants reported experience with other languages, with the age of second language acquisition ranging from 1 to 28 years.

Stimuli

The continuous speech consisted of an audiobook recording of “Northern Lights” (Pullman, 1995) spoken by a female narrator with a southern British English accent. The audiobook was divided into four excerpts presented simultaneously with a 2-Hz isochronous click sequence. All other aspects of stimulus generation were identical to Experiments 1 and 2. The audiobook and click track were each presented in both ears. The same speech features investigated in Experiments 1 and 2 were extracted and analyzed according to the same steps as Experiments 1 and 2.

Procedure

The procedure for Experiment 3 was identical to Experiments 1 and 2.

Data Processing and Analysis

The data processing and analysis protocol was identical to that of Experiments 1 and 2. Of the participants included, the percentage of taps removed on our predefined exclusion criteria ranged from 0.14% to 9.97% (M = 1.81%, SD = 1.70%).

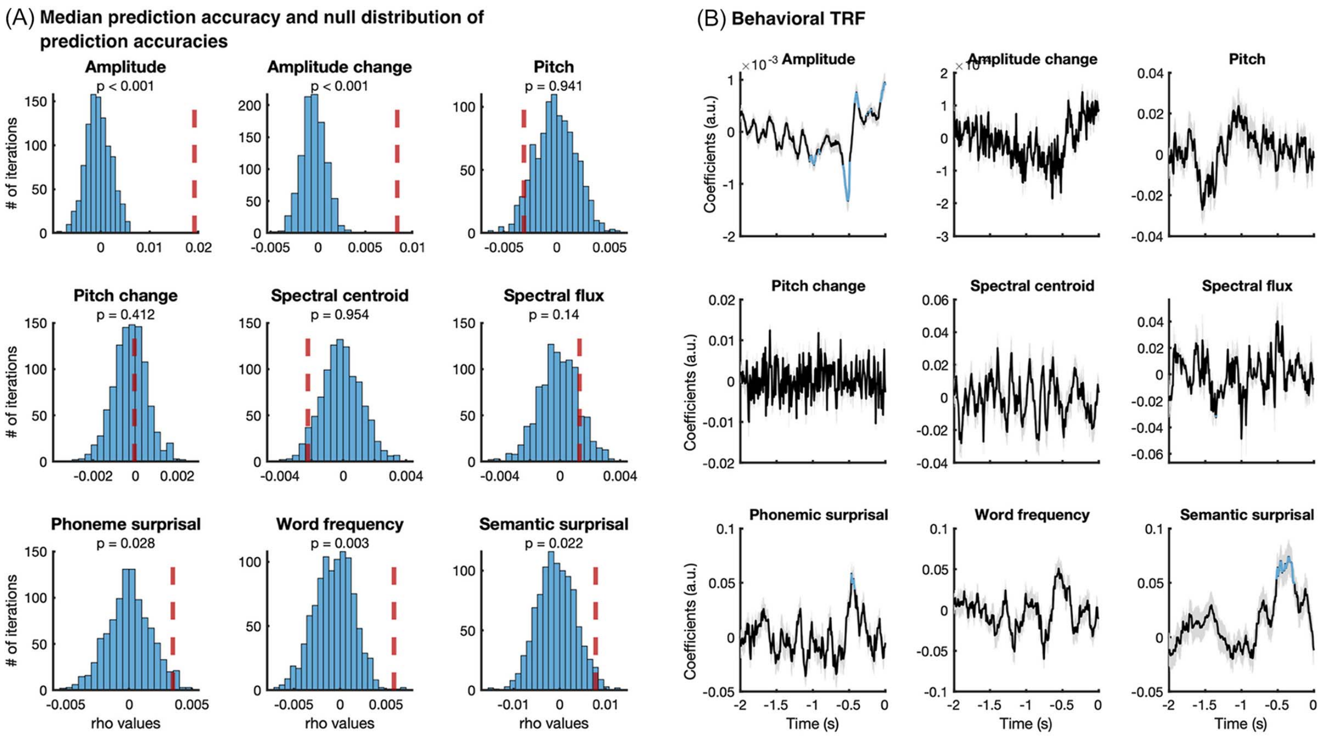
Results

Results using a new audiobook with a different talker show that shifts in tapping asynchrony are linked to amplitude, amplitude change, phonemic surprisal, word frequency, and semantic surprisal (Figure 6). The time course of the effects appears broadly consistent with Experiments 1 and 2, with higher amplitude, greater amplitude change, and phonemic surprisal variations linked to tapping shifts approximately 550 ms later. However, significant effects were not observed for pitch change or spectral flux in this experiment, suggesting that the extent to which these features elicit tapping shifts may be speaker-specific.

As with Experiments 1 and 2, we ran a follow-up analysis covarying out amplitude to test if the relationship between tapping asynchrony and phonemic surprisal was driven by variations in amplitude (Figure 7). Even when covarying for amplitude, tapping asynchrony significantly predicted variations in phonemic surprisal

Figure 6

Experiment 3: Relationship Between Tapping Asynchrony and Features of Task-Irrelevant Speech



Note. Panel A shows the prediction accuracy. The dashed line shows the correlation coefficient (Spearman's rho) representing the relationship between the time series of each speech feature and the predicted time series based on tapping asynchrony as estimated by the multivariate TRF model. Histograms show the permutation-generated null distribution of rho values, representing the relationship between the time series of each speech feature and the time series predicted by the shuffled tapping data. The p values represent the probability of the observed rho value given the distribution of rho values obtained when shuffling the tapping data. Panel B shows the behavioral TRF. Coefficients (in a.u.) represent the degree to which tapping asynchrony predicts the speech feature at each time lag. Along the x-axis, the zero time lag indicates the onset of the click to which participants were attempting to synchronize. Along the y-axis, a positive coefficient indicates that a higher value (e.g., larger amplitude) is associated with later tapping while a negative coefficient indicates that a higher value is associated with earlier tapping. Thick blue lines represent lags at which the coefficients significantly differ from zero (with false discovery rate correction for multiple comparisons). TRF = temporal response function; a.u. = arbitrary units. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

and word frequency. Effects of amplitude change and semantic surprisal were no longer significant after covarying out amplitude.

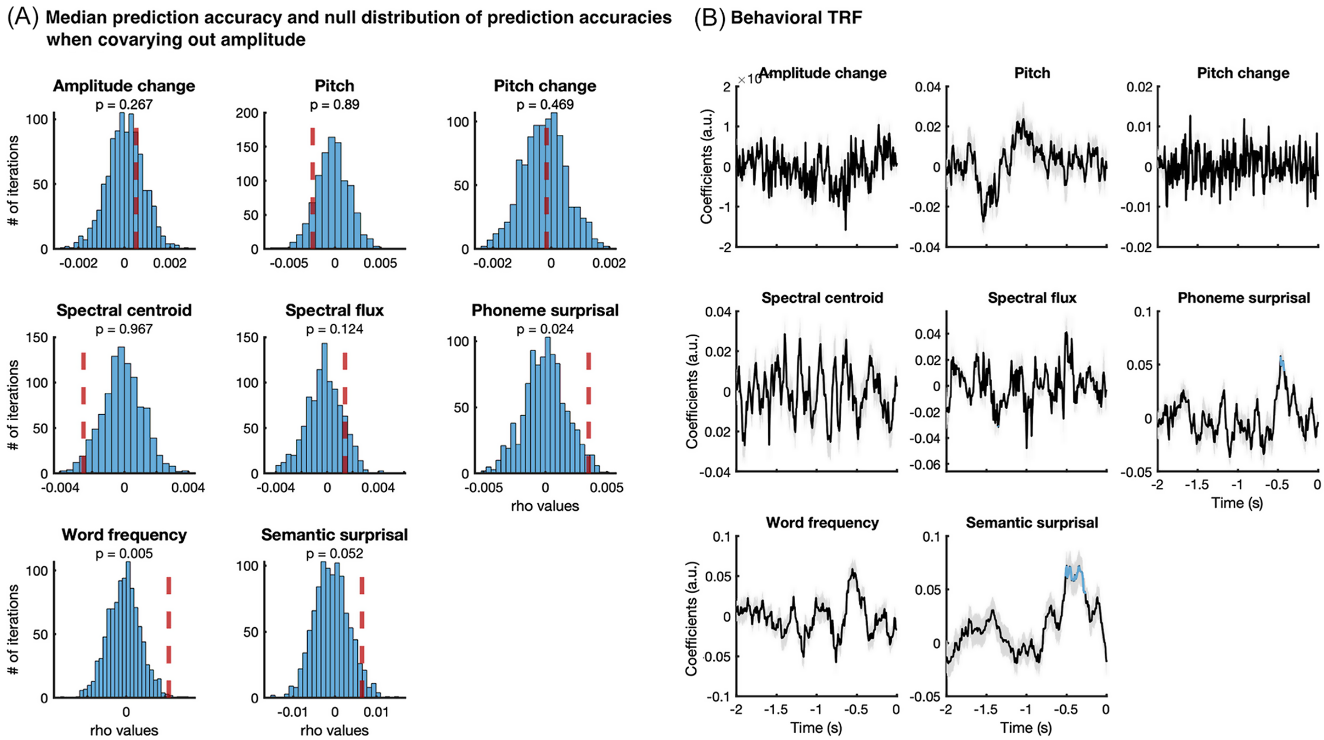
Discussion

With a new speaker and audiobook, Experiment 3 replicates the previous two experiments, showing that salient variations in acoustic and linguistic features lead to earlier tapping. These effects were consistently observed for amplitude measures as well as phoneme surprisal and word frequency. However, contrary to Experiments 1 and 2, tapping shifts were not linked to pitch change or spectral flux. One simple explanation for these null results in Experiment 3 is that the speech excerpts in Experiment 3 contained less pitch and spectral variation. In support of this explanation, Experiment 1 excerpts contained larger pitch changes and greater variation in the spectrum compared to Experiment 3 ($p < .001$; see the Supplemental Materials for further details). In line with our prior work showing no significant tapping shifts following small (one semitone) pitch changes (Symons et al., 2024), it may be that a certain threshold must be reached in order for acoustic variations to elicit the increases in arousal, which may underpin shifts in tapping asynchrony.

Across Experiments 1 through 3, the strongest evidence that linguistic predictability can lead to changes in arousal came from phoneme surprisal, which was significantly related to tapping asynchrony in all three experiments. However, phoneme surprisal only remained significantly related to tapping asynchrony after covarying out amplitude in Experiments 1 and 3, while this analysis was not significant in Experiment 2 (p = .11). Given that, in Experiments 1 and 3, the click and stimuli were located in the same ears, this raises the possibility that acoustic masking or perceptual fusion somehow drove the effect of phoneme surprisal on tapping asynchrony in these experiments. To rule out this possibility, we ran a final, preregistered experiment with the exact same setup as Experiment 1, except that the stimuli and clicks were presented in opposite ears. An additional motivation behind this experiment was, more generally, to determine the robustness of the effects that were significant in some experiments but not others, such as the effects of pitch change and spectral flux. One possible

Figure 7

Experiment 3: Relationship Between Tapping Asynchrony and Features of Task-Irrelevant Speech (Covarying for Amplitude)



Note. Panel A shows the prediction accuracy when covarying out amplitude. Amplitude was covaried out by regressing each speech feature against amplitude and extracting the residuals. The dashed line shows the correlation coefficient (Spearman's rho) representing the relationship between the time series of each speech feature (with amplitude removed) and the predicted time series based on tapping asynchrony as estimated by the multivariate TRF model. Histograms show the permutation-generated null distribution of rho values, representing the relationship between the time series of each speech feature (with amplitude removed) and the time series predicted by the shuffled tapping data. The *p* values represent the probability of the observed rho value given the distribution of rho values obtained when shuffling the tapping data. Panel B shows the behavioral TRF. Coefficients (in arbitrary units) represent the degree to which tapping asynchrony predicts the speech feature after removing the contribution of amplitude at each time lag. Along the *x*-axis, the zero time lag indicates the onset of the click to which participants were attempting to synchronize. Along the *y*-axis, a positive coefficient indicates that a higher residual value (e.g., greater phonemic surprisal after accounting for amplitude) is associated with later tapping while a negative coefficient indicates that a higher value is associated with earlier tapping. Thick blue lines represent lags at which the coefficients significantly differ from zero (with false discovery rate correction for multiple comparisons). TRF = temporal response function; a.u. = arbitrary units. See the online article for the color version of this figure.

explanation for why these effects were significant in some experiments and not others is that these effects may be small relative to overall variability in tapping asynchrony, and that previous experiments may have been underpowered to reliably detect them. To address this possibility, we used power analysis to determine the number of participants necessary to elicit a robust effect of phoneme surprisal; the resulting sample size was substantially larger than that of the other experiments.

Experiment 4

Overview

Our goal for Experiment 4 was to determine whether the effect of phoneme surprisal on tapping asynchrony, after covarying out amplitude, reported in Experiments 1 and 3 would replicate if the speech and clicks were presented in opposite ears. Experiment 4 was an exact replication of Experiment 1, except that the stimuli and clicks were in opposite ears, and the sample size was based on a power analysis of the effect of phoneme surprisal.

Method

Participants

The number of participants recruited for Experiment 4 was based on a power analysis, making use of the effect size for the phoneme surprisal effect (covarying for amplitude) collapsing across the data from Experiments 1 through 3. However, we were unable to make use of our previous Monte Carlo analysis protocol when calculating power, due to computational limitations. Instead, to calculate effect size, we compared rho values across participants to zero using a *t* test. (Our use of a doubly nested TRF design avoids data leakage and overfitting.) Collapsing across all three experiments, the effect size for phoneme surprisal, covarying for amplitude, was $dz = 0.32$. A power calculation in G*Power indicated that we would need 178 participants (after exclusion) to achieve a power of 0.95 with an α of $p < .01$. Due

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

to the likelihood that we would need to exclude a fair portion of our participants due to high tapping variability and other issues, we planned to collect data from 350 participants.

A sample of 351 participants between ages 18 and 40 years (M = 31.5, SD = 5.4; 171 female, 173 male, six nonbinary, one preferred not to report) was recruited from the Prolific (https://www.prolific.com/) recruitment platform. Due to the NIH funding requirements, data on race and ethnicity were collected. We placed no geographic restrictions on Prolific, and therefore, racial and ethnic categories may not have applied to participants outside of the United States. However, we report them here for completeness: From the original sample, 340 participants reported their ethnicity as not Hispanic or Latino, seven reported their ethnicity as Hispanic or Latino, and four preferred not to report. Two participants were American Indian/Alaska Native, two were Asian, 56 participants were Black or African American, 244 were White, 12 were more than one race, and five participants preferred not to report.

Data from participants who did not report English as their native language (N = 24) were excluded from analysis. Additionally, participants who did not tap during one or more run (N = 5), tapped out of phase with the clicks (N = 8), had tapping variability >100 ms (N = 82), showed substantial (>15 ms) keyboard quantization (N = 11), or had fewer than 70% of valid taps (N = 2) were removed.

The final sample consisted of 221 participants between the ages of 19 and 40 years (M = 31.8, SD = 5.3; 111 female, 106 male, four nonbinary). Of this sample, 103 participants reported receiving musical training (ranging from 1 to 16 years). Only 15 could be classified as musicians based on a 6-year criterion (Zhang et al., 2020). Sixty-eight participants reported experience with other languages, with the age of second language acquisition ranging from 1 to 35 years.

Stimuli

The stimuli were identical to those used in Experiment 1, except that the speech and click were presented in opposite ears. The ear in which the speech was presented was counterbalanced across participants.

Procedure

The procedure for Experiment 4 was identical to Experiments 1, 2, and 3.

Data Processing and Analysis

The data processing and analysis protocol was identical to the protocol followed in Experiments 1, 2, and 3.

Transparency and Openness

The experiment design and analysis protocol was preregistered and can be viewed at https://aspredicted.org/9eb3db.pdf.

Results

In Experiment 4, we found that all nine stimulus features, including both acoustic and linguistic features, were significantly related to tapping (Figure 8). The time course of the effects was consistent with Experiments 1, 2, and 3, with greater amplitude, acoustic change, and phonemic and semantic surprisal linked to earlier tapping between 500 and 1,500 ms prior to the tap, with the function peaking just before 500 ms. After covarying out amplitude, pitch change and spectral centroid were no longer significant; however, amplitude change, pitch, spectral flux, phoneme surprisal, word frequency, and semantic surprisal continued to be significantly related to tapping asynchrony (Figure 9).

Discussion

In Experiment 4, we find that even after covarying for amplitude, acoustic change and linguistic surprisal remain significantly related to tapping asynchrony. Given that the tap and the click were presented in different ears in this experiment, this finding rules out an explanation based on acoustic masking effects or perceptual fusion of the click with events in the speech. Instead, these results suggest that stimulus features such as acoustic change and linguistic unpredictability lead to an increase in arousal, expanding time perception and leading participants to briefly speed up their tapping.

Cross-Experiment Comparison

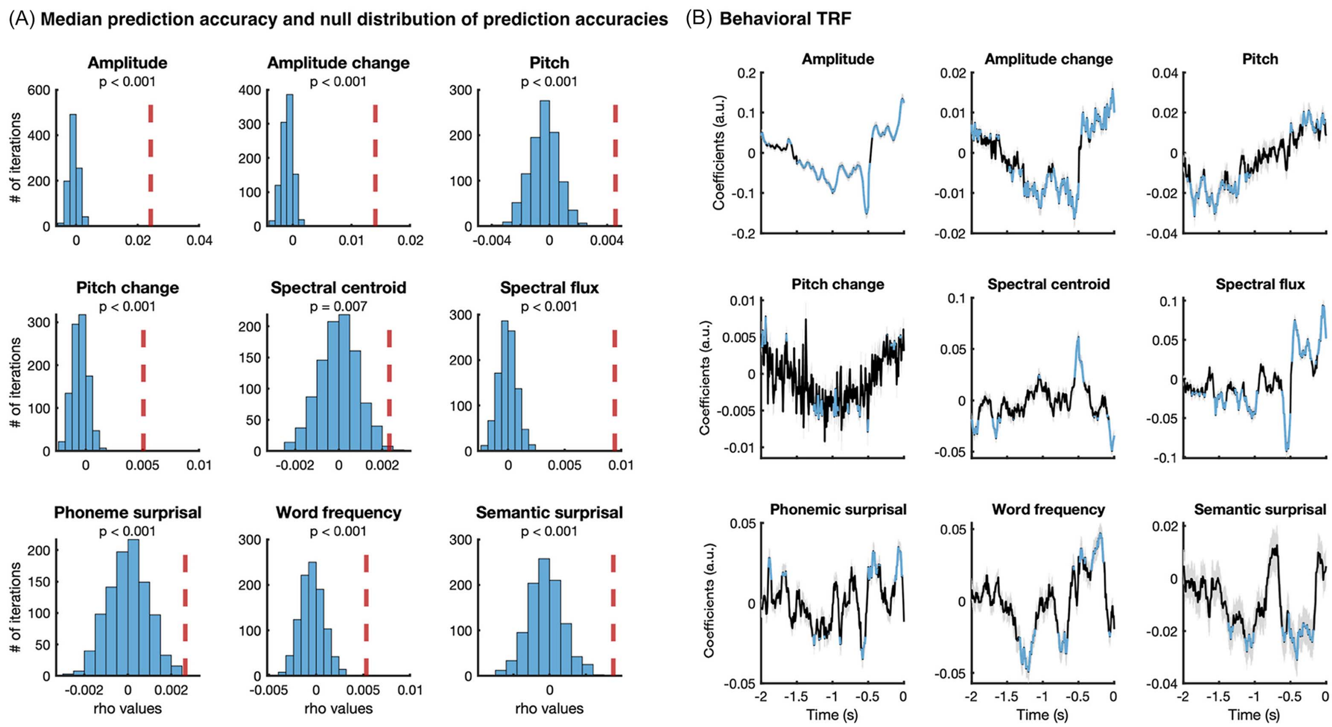
To give the reader an overview of the results, Table 1 shows a summary of the findings from Experiments 1, 2, 3, and 4. Variation in amplitude, amplitude change, and phoneme surprisal were consistently linked to changes in tapping asynchrony across all experiments. Significant pitch change, spectral flux, and semantic surprisal effects on tapping asynchrony were observed in three of four experiments. Pitch and word frequency showed more inconsistent results and were potentially driven by differences in the participant sample or specific excerpts used.

We then used Spearman's correlations to compare the TRF shapes for the three features that showed significant effects across all four experiments (amplitude, amplitude change, and phoneme surprisal). For each feature, p values were corrected for multiple correlations (six tests) using Bonferroni correction. As shown in Figure 10, the shapes of the TRFs were similar across experiments. TRF shape correlations were significant for all three features, with every pairwise comparison between experiments being significant at $p < .001$. For amplitude, the average pairwise correlation between the TRF shapes across experiments was $\rho = 0.91$. For amplitude change, the average pairwise correlation across experiments was 0.82. For phoneme surprisal, the average pairwise correlation across experiments was 0.46.

Overall, our results suggest that the TRF shapes for amplitude and phoneme surprisal features are replicable across different speakers and participant samples. Moreover, the shape of the TRF is somewhat similar, regardless of whether the clicks and the speech are presented in the same ear (Experiments 1 and 3) versus opposite ears (Experiments 2 and 4), suggesting that the link between speech features and tapping speed is for the most part not driven by perceptual fusion. These cross-experiment comparisons suggest that, when speech features are sufficiently salient, the effects on tapping behavior are stable across different participant samples and speech excerpts. For amplitude, amplitude change, and phoneme surprisal measures, the time course of tapping shifts significantly correlated across all four experiments. Taken together, results from all four experiments suggest that acoustic and linguistic features of continuous natural speech can increase arousal and disrupt internal timekeeping. However, the degree of

Figure 8

Experiment 4: Relationship Between Tapping Asynchrony and Features of Task-Irrelevant Speech



Note. Panel A shows the prediction accuracy. The dashed line shows the correlation coefficient (Spearman's rho) representing the relationship between the time series of each speech feature and the predicted time series based on tapping asynchrony as estimated by the multivariate TRF model. Histograms show the permutation-generated null distribution of rho values, representing the relationship between the time series of each speech feature and the time series predicted by the shuffled tapping data. The *p* values represent the probability of the observed rho value given the distribution of rho values obtained when shuffling the tapping data. Panel B shows the behavioral TRF. Coefficients (in arbitrary units) represent the degree to which tapping asynchrony predicts the speech feature at each time lag. Along the *x*-axis, the zero time lag indicates the onset of the click to which participants were attempting to synchronize. Along the *y*-axis, a positive coefficient indicates that a higher value (e.g., larger amplitude) is associated with later tapping while a negative coefficient indicates that a higher value is associated with earlier tapping. Thick blue lines represent lags at which the coefficients significantly differ from zero (with false discovery rate correction for multiple comparisons). TRF = temporal response function; a.u. = arbitrary units. See the online article for the color version of this figure.

behavioral disruption in response to a given speech sample may depend on the amount of acoustic or linguistic variation in the signal.

General Discussion

Across four experiments, we find that as listeners tap to a series of clicks, dynamic features of task-irrelevant speech continuously distort subjective time, causing tap timing to drift forward and backward. This distortion is found even when the speech and clicks are presented in separate ears. The best fitting explanation for this effect is that changes in salience of task-irrelevant speech are linked to changes in arousal, which, in turn, modulate the speed of internal timekeeping processes (Gibbon et al., 1984), biasing listeners' estimates of when the next click will occur. Importantly, we find that tap timing is related not only to acoustic characteristics of speech, most notably amplitude, but also to phonemic surprisal. These results suggest that attentional orienting to sound takes place after the initial stages of linguistic analysis.

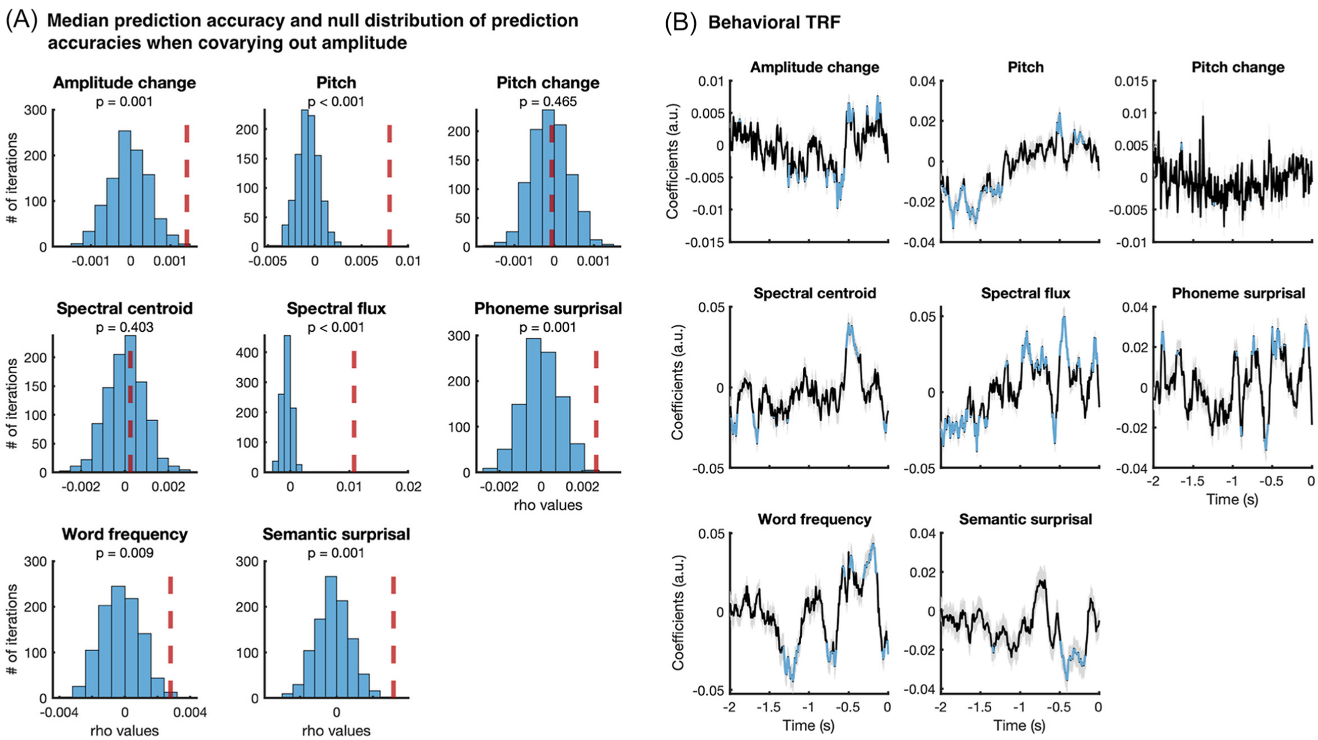
These results are broadly consistent with the predictions made by computational models of the salience of auditory scenes. For example, a common prediction across many models is that sudden changes in acoustic features will capture attention (Duangudom & Anderson, 2007; Kalinli & Narayanan, 2007; Kayser et al., 2005), inspired by vision research using eye-tracking data as ground truth (Niebur et al., 2002), and indeed, we find here that subjective time expands roughly 500 ms after acoustic change. Our results also confirm the role of predictability in salience, as postulated by other models (Kaya & Elhilali, 2014; Tsuchida & Cottrell, 2012). On the other hand, our results suggest that computational models of salience designed to apply to a broad class of auditory scenes may not capture some of the factors driving attentional capture by task-irrelevant speech. In particular, future computational models of the salience of speech should incorporate phonemic and semantic predictability. Future research could also investigate whether there are additional domain-specific factors driving the salience of auditory scenes, such as the predictability of melody (Di Liberto et al., 2020) and rhythm (Zalta et al., 2024) in music.

Our leading explanation for why salient speech distorts synchronization timing is that changes in the salience of ongoing sounds modulate arousal, which in turn expands and contracts subjective time. This explanation fits our finding that the effects of task-irrelevant speech on synchronization are maintained even when the speech and

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Figure 9

Experiment 4: Relationship Between Tapping Asynchrony and Features of Task-Irrelevant Speech (Covarying for Amplitude)



Note. Panel A shows the prediction accuracy when covarying out amplitude. Amplitude was covaried out by regressing each speech feature against amplitude and extracting the residuals. The dashed line shows the correlation coefficient (Spearman's rho) representing the relationship between the time series of each speech feature (with amplitude removed) and the predicted time series based on tapping asynchrony as estimated by the multivariate TRF model. Histograms show the permutation-generated null distribution of rho values, representing the relationship between the time series of each speech feature (with amplitude removed) and the time series predicted by the shuffled tapping data. The *p* values represent the probability of the observed rho value given the distribution of rho values obtained when shuffling the tapping data. Panel B shows the behavioral TRF. Coefficients (in arbitrary units) represent the degree to which tapping asynchrony predicts the speech feature after removing the contribution of amplitude at each time lag. Along the *x*-axis, the zero time lag indicates the onset of the click to which participants were attempting to synchronize. Along the *y*-axis, a positive coefficient indicates that a higher residual value (e.g., greater phoneme surprisal after accounting for amplitude) is associated with later tapping while a negative coefficient indicates that a higher value is associated with earlier tapping. Thick blue lines represent lags at which the coefficients significantly differ from zero (with false discovery rate correction for multiple comparisons). TRF = temporal response function; a.u. = arbitrary units. See the online article for the color version of this figure.

clicks are presented in alternate ears. Direct evidence for this explanation, however, would require concurrent measurement of one or more physiological measures of arousal. While internal timing distortions have been linked to pupil dilation in monkeys (Suzuki et al., 2016), the link between physiological arousal and subjective time remains inconclusive in humans (Williams et al., 2020). Future research, therefore, should measure interference with synchronization by task-irrelevant sounds alongside physiological assessments (e.g., pupil dilation or galvanic skin response) to determine whether physiological arousal and distortion of subjective time correlate across time.

Precise perception of time is vitally important for perceiving speech and music. Differences in voice onset time between voiced and unvoiced consonants, for example, are around 40 ms on average (Morris et al., 2008), and timing also helps convey phrase boundaries (Streeter, 1978), linguistic focus (Ip & Cutler, 2022), and lexical stress (Severijnen et al., 2024). Although music listening tends to place less stringent demands upon temporal processing than speech listening (Albouy et al., 2020), during performance, musicians must synchronize with each other by rapidly correcting differences in timing between movement and sound that can be as small as 3 ms (Madison & Merker, 2004; B. H. Repp, 2000). Given the importance of precise perception and production of timing for human communication, it is surprising and somewhat disconcerting that time perception is constantly subject to distortion by task-irrelevant sound. However, the effect sizes we report here and in Symons et al. (2024) are small enough that these distortions are unlikely to affect conscious perception. In Symons et al., for example, we reported that the sudden onset of a highly aggravating drill sound distorted tap timing by only 4–5 ms. This is well below the average threshold for conscious perception of differences in interval timing, which is around 20 ms (Madison & Merker, 2004). Here, we show that the relationship between speech salience and temporal distortion is relatively weak, with mean correlations between predicted and actual amplitude of around 0.02. Most of the variability in participants' tapping, therefore, was driven by other factors, such as intrinsic motor and timekeeper variability, as well as the accuracy of auditory–motor error correction (Krause et al., 2010;

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Table 1
Summary of Experiment 1 Through 4 Results

	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
Stimulus feature	Main	Covary amp	Main	Covary amp	Main	Covary amp	Main	Covary amp
Amplitude	***		***		***		***	
Amplitude change	***	.	***	*	***	.	***	***
Pitch	.	.	*	***	.	.	***	***
Pitch change	***	.	***	**	.	.	***	.
Spectral centroid	.	.	.	.	.	.	**	.
Spectral flux	***	***	***	***	.	.	***	***
Phoneme surprisal	*	*	**	.	*	*	***	**
Word frequency	.	.	.	.	**	**	***	**
Semantic surprisal	.	.	***	***	*	.	***	***

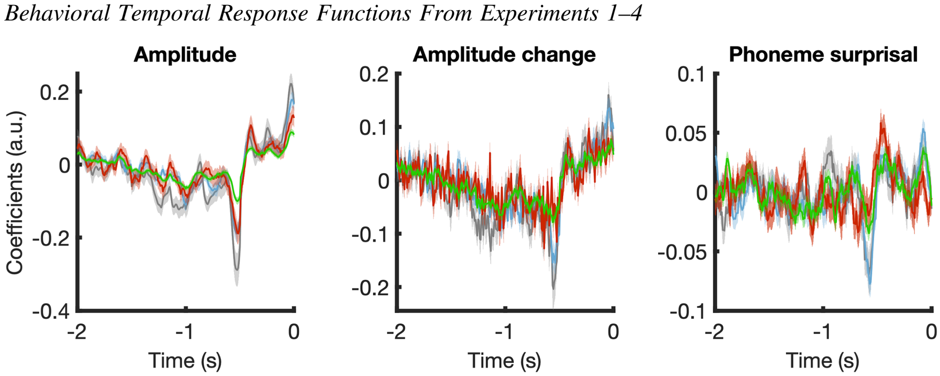
Note. Cells with periods are not significant. amp = amplitude.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Semjen et al., 1998; Wing & Kristofferson, 1973). However, there are reliable individual differences in the extent to which timing is distorted by the salience of task-irrelevant sounds (Symons et al., 2024), with a few participants' tapping distorted by as much as 20 ms, an effect potentially strong enough to interfere with conscious perception of timing. Future research could investigate whether fluctuations in the salience of naturalistic sound streams can have meaningful impacts on the ability to carry out perceptual-motor tasks requiring fine temporal precision. For example, analysis of live jazz performances could investigate whether the acoustic and statistical characteristics of improvised solos are linked to small distortions in the timing of the rhythm section.

The paradigm we introduce here for assessing attentional orienting to task-irrelevant speech has several advantages, making it a useful tool for answering outstanding questions about auditory salience. First, it is highly reliable, with cross-condition correlations of up to $r = 0.88$ (Symons et al., 2024). Second, as demonstrated here, it can be used to simultaneously investigate the role of many different attributes of complex, naturalistic sound signals in driving attentional capture. Third, it can be conducted online and in a relatively short amount of time (total experiment length was 20 min on average). Finally, synchronization is a task that measures a natural behavior (Savage et al., 2015), is simple to explain, and is accessible to almost any experimental population, including children as young as 3 years old (Kirschner & Tomasello, 2009; Woodruff Carr et al., 2014). Nevertheless, because it does not rely on responses being correct or incorrect, even typically developing adults do not perform at ceiling (Thompson et al., 2015).

One weakness of these studies is that they were conducted online, and therefore, we cannot say anything definite about the environment in which the participants carried out the experiments. Some participants, for example, may have completed the experiments in a distracting listening environment, which could lessen the effect of task-irrelevant speech—participants who are already distracted by fluctuations in

Figure 10
Behavioral Temporal Response Functions From Experiments 1–4



Note. The figure shows the behavioral temporal response functions from Experiment 1 (clicks and speech in the same ear, blue line), Experiment 2 (clicks and speech in opposite ears, black line), Experiment 3 (different speaker and audiobook in the same ear, red line), and Experiment 4 (clicks and speech in opposite ears, green line). Coefficients (in a.u.) represent the degree to which tapping asynchrony predicts the speech feature at each time lag, with positive coefficients indicating a higher value (e.g., larger amplitude) is associated with later tapping and negative coefficients indicating that a higher value is associated with earlier tapping. The zero time lag indicates the onset of the click to which participants were attempting to synchronize. For visualization purposes only, amplitude measures were rescaled by dividing coefficients by the maximum value. a.u. = arbitrary units. See the online article for the color version of this figure.

ambient background noise may be less affected by the addition of yet another sound stream. Our study may, therefore, underestimate the effects of task-irrelevant speech on subjective time, and the null effects we report here (such as the lack of influence of semantic surprisal in Experiment 1) should be interpreted with caution. Nevertheless, our prior study (Symons et al., 2024) using a simpler version of the paradigm in which distracting sounds and sound events occurred between clicks found highly similar results when comparing in-lab versus online participants, suggesting that our results here are likely to generalize to the laboratory.

In summary, we use a synchronization paradigm to show that subjective time is destabilized by the presence of distracting naturalistic speech, constantly speeding up and slowing down in response to fluctuations in speech salience. These distortions are driven by both acoustic and linguistic factors in task-irrelevant speech, suggesting that attentional orienting takes place after at least the early stages of linguistic analysis.

Constraints on Generality

All participants were native speakers of English. It remains an open question, therefore, whether individuals are more or less distracted by speech in their native language compared to a less familiar language. This question could have important practical consequences for immigrants attempting to concentrate and resist distraction in, for example, university classrooms, but prior research on this topic has not reached a consensus. While some research has found that disruption of visual serial memory is greater when participants are exposed to their native language as opposed to an unfamiliar language (Ellermeier & Zimmer, 2014; Ellermeier et al., 2015), this effect is relatively small and has not been consistently replicated (Komar et al., 2024). Moreover, the underlying mechanisms remain unclear, potentially reflecting either attentional orienting or interference with preconscious processes (Hughes, 2014). On the other hand, electroencephalography research has found instead that cortical tracking of acoustic features of speech is greater for nonnative compared to native listeners (Reetzke et al., 2021; Zou et al., 2019). One possible explanation of these conflicting results is that linguistic familiarity and proficiency modulate distraction by different speech features in different ways, which could be tested using the current paradigm.

References

Albouy, P., Benjamin, L., Morillon, B., & Zatorre, R. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science*, *367*(6481), 1043–1047. https://doi.org/10.1126/science.aaz3468

Allman, M., & Meck, W. (2012). Pathophysiological distortions in time perception and timed performance. *Brain*, *135*(3), 656–677. https://doi.org/10.1093/brain/awr210

Allman, M., Teki, S., Griffiths, T., & Meck, W. (2014). Properties of the internal clock; first- and second-order principles of subjective time. *Annual Review of Psychology*, *65*, 743–771. https://doi.org/10.1146/annurev-psych-010213-115117

Anderson, A. J., Davis, C., & Lalor, E. C. (2024). Deep-learning models reveal how context and listener attention shape electrophysiological correlates of speech-to-language transformation. *PLOS Computational Biology*, *20*(11), Article e1012537. https://doi.org/10.1371/journal.pcbi.1012537

Antikainen, J., & Niemi, P. (1983). Neuroticism and the pupillary response to a brief exposure to noise. *Biological Psychology*, *17*(2–3), 131–135. https://doi.org/10.1016/0301-0511(83)90013-3

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Bala, A., & Takahashi, T. (2000). Pupillary dilation response as an indicator of auditory discrimination in the barn owl. *Journal of Comparative Physiology*, *186*(5), 425–434. https://doi.org/10.1007/s003590050442

Barry, R. (1975). Low-intensity auditory stimulation and the GSR orienting response. *Physiological Psychology*, *3*(1), 98–100. https://doi.org/10.3758/BF03326832

Bell, R., Röer, J., Dentale, S., & Buchner, A. (2012). Habituation of the irrelevant sound effect: Evidence for an attentional theory of short-term memory disruption. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1542–1557. https://doi.org/10.1037/a0028459

Bell, R., Röer, J., Lang, A., & Buchner, A. (2019). Distraction by steady-state sounds: Evidence for a graded attentional model of auditory distraction. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(4), 500–512. https://doi.org/10.1037/xhp0000623

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berens, P. (2009). CircStat: A MATLAB toolbox for circular statistics. *Journal of Statistical Software*, *31*(10), 1–21. https://doi.org/10.18637/jss.v031.i10

Berti, S., Roeber, U., & Schröger, E. (2004). Bottom-up influences on working memory: Behavioral and electrophysiological distraction varies with distractor strength. *Experimental Psychology*, *51*(4), 249–257. https://doi.org/10.1027/1618-3169.51.4.249

Bonmassar, C., Scharf, F., Widmann, A., & Wetzel, N. (2023). On the relationship of arousal and attentional distraction by emotional novel sounds. *Cognition*, *237*, Article 105470. https://doi.org/10.1016/j.cognition.2023.105470

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, *28*(24), 3976–3983. https://doi.org/10.1016/j.cub.2018.10.042

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803–809.e3. https://doi.org/10.1016/j.cub.2018.01.080

Broderick, M. P., Zuk, N. J., Anderson, A. J., & Lalor, E. C. (2022). More than words: Neurophysiological correlates of semantic dissimilarity depend on comprehension of the speech narrative. *European Journal of Neuroscience*, *56*(8), 5201–5214. https://doi.org/10.1111/ejn.15805

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Buffardi, L. (1971). Factors affecting the filled-duration illusion in the auditory, tactual, and visual modalities. *Perception & Psychophysics*, *10*(4), 292–294. https://doi.org/10.3758/BF03212828

Chen, Y., Repp, B., & Patel, A. (2002). Spectral decomposition of variability in synchronization and continuation tapping: Comparisons between auditory and visual pacing and feedback conditions. *Human Movement Science*, *21*(4), 515–532. https://doi.org/10.1016/S0167-9457(02)00138-0

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, Article 604. https://doi.org/10.3389/fnhum.2016.00604

Di Liberto, G. M., O'sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*(19), 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Di Liberto, G. M., Pelofi, C., Bianco, R., Patel, P., Mehta, A. D., Herrero, J. L., de Cheveigné, A., Shamma, S., & Mesgarani, N. (2020). Cortical encoding of melodic expectations in human temporal cortex. *eLife*, *9*, Article e51784. https://doi.org/10.7554/eLife.51784

Dorsi, J., Viswanathan, N., Rosenblum, L., & Dias, J. (2018). The role of speech fidelity in the irrelevant sound effect: Insights from noise-vocoded speech backgrounds. *Quarterly Journal of Experimental Psychology*, *71*(10), 2152–2161. https://doi.org/10.1177/1747021817739257

Droit-Volet, S., Brunot, S., & Niedenthal, P. (2004). Perception of the duration of emotional events. *Cognition and Emotion*, *18*, 849–858. https://doi.org/10.1080/02699930341000194

Droit-Volet, S., & Meck, W. (2007). How emotions colour our perception of time. *Trends in Cognitive Sciences*, *11*(12), 504–513. https://doi.org/10.1016/j.tics.2007.09.008

Droit-Volet, S., & Wearden, J. (2002). Speeding up an internal clock in children? Effects of visual flicker on subjective duration. *Quarterly Journal of Experimental Psychology*, *55*(3B), 193–211. https://doi.org/10.1080/02724990143000252

Duangudom, V., & Anderson, D. (2007). *Using auditory saliency to understand complex auditory scenes* [Conference session]. 15th European Signal Processing Conference, Poznan, Poland. https://ieeexplore.ieee.org/document/7098996

Ellermeier, W., Kattner, F., Ueda, K., Duomoto, K., & Nakajima, Y. (2015). Memory disruption by irrelevant noise-vocoded speech: Effects of native language and the number of frequency bands. *Journal of the Acoustical Society of America*, *138*(3), 1561–1569. https://doi.org/10.1121/1.4928954

Ellermeier, W., & Zimmer, K. (2014). The psychoacoustics of the irrelevant sound effect. *Acoustical Science & Technology*, *35*(1), 10–16. https://doi.org/10.1250/ast.35.10

Friedman, D., Hakerem, G., Sutton, S., & Fleiss, J. (1973). Effect of stimulus uncertainty on the pupillary dilation response and the vertex evoked potential. *Electroencephalography and Clinical Neurophysiology*, *34*(5), 475–484. https://doi.org/10.1016/0013-4694(73)90065-5

Gibbon, J., Church, R., & Meck, W. (1984). Scalar timing in memory. *Annals of the New York Academy of Sciences*, *423*(1), 52–77. https://doi.org/10.1111/j.1749-6632.1984.tb23417.x

Gil, S., & Droit-Volet, S. (2012). Emotional time distortions: The fundamental role of arousal. *Cognition and Emotion*, *26*(5), 847–862. https://doi.org/10.1080/02699931.2011.625401

Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T., & Brodbeck, C. (2021). Neural markers of speech comprehension: Measuring EEG tracking of linguistic speech representations, controlling the speech acoustics. *Journal of Neuroscience*, *41*(50), 10316–10329. https://doi.org/10.1523/JNEUROSCI.0812-21.2021

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, *39*(3), 192–193. https://jcaa.caa-aca.ca/index.php/jcaa/article/view/2476

Huang, N., & Elhilali, M. (2017). Auditory salience using natural soundscapes. *Journal of the Acoustical Society of America*, *141*(3), 2163–2176. https://doi.org/10.1121/1.4979055

Huang, N., & Elhilali, M. (2020). Push–pull competition between bottom-up and top-down auditory attention to natural soundscapes. *eLife*, *9*, Article e52984. https://doi.org/10.7554/eLife.52984

Hughes, R. (2014). Auditory distraction: A duplex-mechanism account. *PsyCh Journal*, *3*(1), 30–41. https://doi.org/10.1002/pchj.44

Ip, M. H. K., & Cutler, A. (2022). In search of salience: Focus detection in the speech of different talkers. *Language and Speech*, *65*(3), 650–680. https://doi.org/10.1177/00238309211046029

Jones, D., Alford, D., Macken, W., Banbury, S., & Tremblay, S. (2000). Interference from degraded auditory stimuli: Linear effects of changing-state in the irrelevant sequence. *Journal of the Acoustical Society of America*, *108*(3), 1082–1088. https://doi.org/10.1121/1.1288412

Jones, D., & Macken, W. (1993). Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 369–381. https://doi.org/10.1037/0278-7393.19.2.369

Kalinli, O., & Narayanan, S. S. (2007). A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. *Interspeech* (Vol. *2007*, pp. 1941–1944). https://doi.org/10.21437/Interspeech.2007

Kaya, E., & Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience*, *8*, Article 327. https://doi.org/10.3389/fnhum.2014.00327

Kaya, E., Huang, N., & Elhilali, M. (2020). Pitch, timbre and intensity interdependently modulate neural responses to salient sounds. *Neuroscience*, *440*, 1–14. https://doi.org/10.1016/j.neuroscience.2020.05.018

Kayser, C., Petkov, C., Lippert, M., & Logothetis, N. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, *15*(21), 1943–1947. https://doi.org/10.1016/j.cub.2005.09.040

Kirschner, S., & Tomasello, M. (2009). Joint drumming: Social context facilitates synchronization in preschool children. *Journal of Experimental Child Psychology*, *102*(3), 299–314. https://doi.org/10.1016/j.jecp.2008.07.005

Komar, G. F., Buchner, A., Mieth, L., Van de Vijver, R., & Bell, R. (2024). Evidence of a metacognitive illusion in stimulus-specific prospective judgments of distraction by background speech. *Scientific Reports*, *14*(1), Article 24111. https://doi.org/10.1038/s41598-024-74719-4

Kothinti, S., & Elhilali, M. (2023). Are acoustics enough? Semantic effects on auditory salience in natural scenes. *Frontiers in Psychology*, *14*, Article 1276237. https://doi.org/10.3389/fpsyg.2023.1276237

Krause, V., Pollok, B., & Schnitzler, A. (2010). Perception in action: The impact of sensory information on sensorimotor synchronization in musicians and non-musicians. *Acta Psychologica*, *133*(1), 28–37. https://doi.org/10.1016/j.actpsy.2009.08.003

Lake, J., LeBar, K., & Meck, W. (2016). Emotional modulation of interval timing and time perception. *Neuroscience and Biobehavioral Reviews*, *64*, 403–420. https://doi.org/10.1016/j.neubiorev.2016.03.003

Le Compte, D., Neely, C., & Wilson, J. (1997). Irrelevant speech and irrelevant tones: The relative importance of speech to the irrelevant speech effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(2), 472–483. https://doi.org/10.1037/0278-7393.23.2.472

Liao, H., Kidani, S., Yoneya, M., Kashino, M., & Furukawa, S. (2016). Correspondences among pupillary dilation response, subjective salience of sounds, and loudness. *Psychonomic Bulletin & Review*, *23*(2), 412–425. https://doi.org/10.3758/s13423-015-0898-0

Liao, H., Yoneya, M., Kashino, M., & Furukawa, S. (2018). Pupillary dilation response reflects surprising moments in music. *Journal of Eye Movement Research*, *11*(2), 1–13. https://doi.org/10.16910/jemr.11.2.13

Little, J., Martin, F., & Thomson, R. (2010). Speech versus non-speech as irrelevant sound: Controlling acoustic variation. *Biological Psychology*, *85*(1), 62–70. https://doi.org/10.1016/j.biopsycho.2010.05.004

Madison, G., & Merker, B. (2004). Human sensorimotor tracking of continuous subliminal deviations from isochrony. *Neuroscience Letters*, *370*(1), 69–73. https://doi.org/10.1016/j.neulet.2004.07.094

Maricq, A., Roberts, S., & Church, R. (1981). Methamphetamine and time estimation. *Journal of Experimental Psychology: Animal Behavior Processes*, *7*(1), 18–30. https://doi.org/10.1037/0097-7403.7.1.18

Marois, A., Labonté, K., Parent, M., & Vachon, F. (2018). Eyes have ears: Indexing the orienting response to sound using pupillometry. *International Journal of Psychophysiology*, *123*, 152–162. https://doi.org/10.1016/j.ijpsycho.2017.09.016

Masson, R., & Bidet-Caulet, A. (2019). Fronto-central P3a to distracting sounds: An index of their arousing properties. *NeuroImage*, *185*, 164–180. https://doi.org/10.1016/j.neuroimage.2018.10.041

Merchant, H., Harrington, D., & Meck, W. (2013). Neural basis of the perception and estimation of time. *Annual Review of Neuroscience*, *36*, 313–336. https://doi.org/10.1146/annurev-neuro-062012-170349

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

pitch. *Behavior Research Methods*, *53*(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

Milne, A. E., Zhao, S., Tampakaki, C., Bury, G., & Chait, M. (2021). Sustained pupil responses are modulated by predictability of auditory sequences. *Journal of Neuroscience*, *41*(28), 6116–6127. https://doi.org/10.1523/JNEUROSCI.2879-20.2021

Mittal, J., Juneja, K. K., Saumya, S., & Shukla, A. (2024). A matter of time: How musical training affects time perception. *Frontiers in Neuroscience*, *18*, Article 1364504. https://doi.org/10.3389/fnins.2024.1364504

Morris, R. J., McCrea, C. R., & Herring, K. D. (2008). Voice onset time differences between adult males and females: Isolated syllables. *Journal of Phonetics*, *36*(2), 308–317. https://doi.org/10.1016/j.wocn.2007.06.003

Niebur, E., Itti, L., & Koch, C. (2002). Controlling the focus of visual selective attention. In J. Cowan, E. Domany, & J. van Hemmen (Eds.), *Models of neural networks IV* (pp. 247–276). Springer. https://doi.org/10.1007/978-0-387-21703-1_6

Ortega, L., & López, F. (2008). Effects of visual flicker on subjective time in a temporal bisection task. *Behavioural Processes*, *78*(3), 380–386. https://doi.org/10.1016/j.beproc.2008.02.004

Penton-Voak, I., Edwards, H., Percival, A., & Wearden, J. (1996). Speeding up an internal clock in humans? Effects of click trains on subjective duration. *Journal of Experimental Psychology: Animal Behavior Processes*, *22*(3), 307–320. https://doi.org/10.1037/0097-7403.22.3.307

Pullman, P. (1995). *Northern lights*. Scholastic Children's Books.

Qiyuan, J., Richer, F., Wgoner, B., & Beatty, J. (1985). The pupil and stimulus probability. *Psychophysiology*, *22*(5), 530–534. https://doi.org/10.1111/j.1469-8986.1985.tb01645.x

Reetzke, R., Gnanateja, G. N., & Chandrasekaran, B. (2021). Neural tracking of the speech envelope is differentially modulated by attention and language experience. *Brain and Language*, *213*, Article 104891. https://doi.org/10.1016/j.bandl.2020.104891

Repp, B. (2004). On the nature of phase attraction in sensorimotor synchronization with interleaved auditory sequences. *Human Movement Science*, *23*, 389–413. https://doi.org/10.1016/j.humov.2004.08.014

Repp, B. H. (2000). Compensation for subliminal timing perturbations in perceptual-motor synchronization. *Psychological Research*, *63*(2), 106–128. https://doi.org/10.1007/PL00008170

Rinne, T., Särkkä, A., Degerman, A., Schröger, E., & Alho, K. (2006). Two separate mechanisms underlie auditory change detection and involuntary control of attention. *Brain Research*, *1077*(1), 135–143. https://doi.org/10.1016/j.brainres.2006.01.043

Röer, J. P., Buchner, A., & Bell, R. (2019). Auditory distraction in short-term memory: Stable effects of semantic mismatches on serial recall. *Auditory Perception & Cognition*, *2*(3), 143–162. https://doi.org/10.1080/25742442.2020.1722560

Röer, J. P., & Cowan, N. (2021). A preregistered replication and extension of the cocktail party phenomenon: One's name captures attention, unexpected words do not. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(2), 234–242. https://doi.org/10.1037/xlm0000874

Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, *112*(29), 8987–8992. https://doi.org/10.1073/pnas.1414495112

Schlittmeier, S., Weißgerber, T., Kerber, S., Fastl, H., & Hellbrück, J. (2012). Algorithmic modelling of the irrelevant sound effect (ISE) by the hearing sensation fluctuation strength. *Attention, Perception, & Psychophysics*, *74*(1), 194–203. https://doi.org/10.3758/s13414-011-0230-7

Schultz, B., Brown, R., & Kotz, S. (2021). Dynamic acoustic salience evoked motor responses. *Cortex*, *134*, 320–332. https://doi.org/10.1016/j.cortex.2020.10.019

Schwarz, M., Winkler, I., & Sedlmeier, P. (2013). The heart beat does not make us tick: The impact of heart rate and arousal on time perception. *Attention, Perception, & Psychophysics*, *75*(1), 182–193. https://doi.org/10.3758/s13414-012-0387-8

Semjen, A., Vorberg, D., & Schulze, H.-H. (1998). Getting synchronized with the metronome: Comparisons between phase and period correction. *Psychological Research*, *61*(1), 44–55. https://doi.org/10.1007/s004260050012

Severijnen, G. G., Bosker, H. R., & McQueen, J. M. (2024). Your "VOORnaam" is not my "VOORnaam": An acoustic analysis of individual talker differences in word stress in Dutch. *Journal of Phonetics*, *103*, Article 101296. https://doi.org/10.1016/j.wocn.2024.101296

Shalev, N., & Nobre, A. (2022). Eyes wide open: Regulation of arousal by temporal expectations. *Cognition*, *224*, Article 105062. https://doi.org/10.1016/j.cognition.2022.105062

Sokolov, E. (1963). Higher nervous functions: The orienting reflex. *Annual Review of Physiology*, *25*, 545–580. https://doi.org/10.1146/annurev.ph.25.030163.002553

Southwell, R., Baumann, A., Gal, C., Barascud, N., Friston, K., & Chait, M. (2017). Is predictability salient? A study of attentional capture by auditory patterns. *Philosophical Transactions of the Royal Society B*, *372*(1714), Article 20160105. https://doi.org/10.1098/rstb.2016.0105

Streeter, L. A. (1978). Acoustic determinants of phrase boundary perception. *The Journal of the Acoustical Society of America*, *64*(6), 1582–1592. https://doi.org/10.1121/1.382142

Suzuki, T., Kunimatsu, J., & Tanaka, M. (2016). Correlation between pupil size and subjective passage of time in non-human primates. *Journal of Neuroscience*, *36*(44), 11331–11337. https://doi.org/10.1523/JNEUROSCI.2533-16.2016

Symons, A. E., Dick, F., Holt, L., & Tierney, A. (2026). *Salient features of task-irrelevant continuous speech distort subjective time*. OSF. https://osf.io/pc5tw

Symons, A. E., Dick, F., & Tierney, A. (2024). Salient sounds distort time perception and production. *Psychonomic Bulletin & Review*, *31*(1), 137–147. https://doi.org/10.3758/s13423-023-02305-2

Teoh, E. S., Cappelloni, M. S., & Lalor, E. C. (2019). Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features. *European Journal of Neuroscience*, *50*(11), 3831–3842. https://doi.org/10.1111/ejn.14510

Thompson, E. C., White-Schwoch, T., Tierney, A., & Kraus, N. (2015). Beat synchronization across the lifespan: Intersection of development and musical experience. *PLOS ONE*, *10*(6), Article e0128839. https://doi.org/10.1371/journal.pone.0128839

Tsuchida, T., & Cottrell, G. (2012). Auditory saliency using natural statistics. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*, 1048–1053. https://escholarship.org/uc/item/30k404zz

Viswanathan, N., Dorsi, J., & George, S. (2014). The role of speech-specific properties of the background in the irrelevant sound effect. *The Quarterly Journal of Experimental Psychology*, *67*(3), 581–589. https://doi.org/10.1080/17470218.2013.821708

Wearden, J. H., Norton, R., Martin, S., & Montford-Bebb, O. (2007). Internal clock processes and the filled-duration illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(3), 716–729. https://doi.org/10.1037/0096-1523.33.3.716

Wearden, J. H., & Penton-Voak, I. (1995). Feeling the heat: Body temperature and the rate of subjective time, revisited. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *48*, 129–141. https://doi.org/10.1080/14640749508401443

Wearden, J. H., Philpott, K., & Win, T. (1999). Speeding up and (… relatively…) slowing down an internal clock in humans. *Behavioural Processes*, *46*(1), 63–73. https://doi.org/10.1016/S0376-6357(99)00004-2

Weissbart, H., Kandylaki, K. D., & Reichenbach, T. (2020). Cortical tracking of surprisal during continuous speech comprehension. *Journal of Cognitive Neuroscience*, *32*(1), 155–166. https://doi.org/10.1162/jocn_a_01467

Williams, E., Solodow, E., Henderson, J., Stewart, A., & Jones, L. (2020). *Do click trains dilate time perception due to physiological arousal?* PsyArXiv. https://doi.org/10.31234/osf.io/78w43

Wing, A. M., & Kristofferson, A. B. (1973). Response delays and the timing of discrete motor responses. *Perception & Psychophysics*, *14*(1), 5–12. https://doi.org/10.3758/BF03198607

Woodruff Carr, K., White-Schwoch, T., Tierney, A. T., Strait, D. L., & Kraus, N. (2014). Beat synchronization predicts neural speech encoding and reading readiness in preschoolers. *Proceedings of the National Academy of Sciences*, *111*(40), 14559–14564. https://doi.org/10.1073/pnas.1406219111

Zalta, A., Large, E. W., Schön, D., & Morillon, B. (2024). Neural dynamics of predictive timing and motor engagement in music listening. *Science Advances*, *10*(10), Article eadi2525. https://doi.org/10.1126/sciadv.adi2525

Zhang, J. D., Susino, M., McPherson, G. E., & Schubert, E. (2020). The definition of a musician in music psychology: A literature review and the six-year rule. *Psychology of Music*, *48*(3), 389–409. https://doi.org/10.1177/0305735618804038

Zhao, S., Chait, M., Dick, F., Dayan, P., Furukawa, S., & Liao, H. I. (2019). Pupil-linked phasic arousal evoked by violation but not emergence of regularity within rapid sound sequences. *Nature Communications*, *10*(1), Article 4030. https://doi.org/10.1038/s41467-019-12048-1

Zhao, S., Yum, N., Bengamin, L., Benhamou, E., Yoneya, M., Furukawa, S., Dick, F., Slaney, M., & Chait, M. (2019). Rapid ocular responses are modulated by bottom-up-driven auditory salience. *Journal of Neuroscience*, *39*(39), 7703–7714. https://doi.org/10.1523/JNEUROSCI.0776-19.2019

Zou, J., Feng, J., Xu, T., Jin, P., Luo, C., Zhang, J., Pan, X., Chen, F., & Ding, N. (2019). Auditory and language contributions to neural encoding of speech features in noisy environments. *NeuroImage*, *192*, 66–75. https://doi.org/10.1016/j.neuroimage.2019.02.047

Received February 4, 2025
Revision received March 20, 2026
Accepted March 30, 2026 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.