How Helpful is Visual Context for Speech Processing? Evidence from Multi-modal Speech Tracking in Monolingual and Bilingual Speakers

Haoyin Xu (hyx002@ucsd.edu), Seana Coulson (scoulson@ucsd.edu)
Department of Cognitive Science, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093, USA

Abstract

Visual cues like facial expressions and gestures enhance speech comprehension. While prior studies have explored L1 multimodal speech processing, research on bilinguals remains limited. Here we examine how visual context influences speech processing in monolinguals and bilinguals by assessing changes in sensitivity to different speech dimensions. EEG was recorded from 24 monolinguals and 24 bilinguals as they viewed multimodal speech clips presented in audiovisual or audio-only formats. Then, a temporal response function was applied to decode neural responses to audio envelope and surprisal to index sensitivity of acoustic and semantic information. Results show that visual context facilitated audio tracking in bilinguals but did not enhance surprisal tracking. Conversely, monolinguals benefited from visual input for surprisal tracking but not envelope tracking. These results suggest that bilinguals may allocate more cognitive resources to audio processing when integrating visual cues, potentially limiting the availability of resources for higher-level semantic processing.

Keywords: Multi-modal Speech Processing, Bilingualism, EEG, Speech Envelope Tracking, Lexical Surprisal Tracking

Introduction

Approximately 50% of the world population is bilingual (Eberhard, Simons, & Fennig, 2023). As international migration increases, many individuals acquire a second language (L2) after their first language (L1). However, despite the fact that language is predominantly used in face-to-face interactions rich in multimodal cues, research on multi-modal comprehension in L2 remains scarce (Akker & Cutler, 2003; Birulés, Bosch, Pons, & Lewkowicz, 2020; Dahl & Ludvigsen, 2014). L2 speakers, particularly those who acquire the language later in life, often face greater processing demands than L1 speakers due to their more limited language experience (Rosenberg, Hirschberg, & Manis, 2010). This increased cognitive effort may reduce their capacity to process non-verbal communicative cues(Lee, Perdomo, & Kaan, 2019). They may also be less familiar with how visual cues contribute to communication in their L2. For instance, linguistic differences between stress-based and tonal languages can influence the perception of emphatic cues, while cultural variation in co-speech gestures may affect comprehension. These challenges might lead L2 speakers to overlook some multimodal cues, though they may also selectively attend to those that aid comprehension, such as specific gestures or tonal features (Dahl & Ludvigsen, 2014; Lin, 2021). McGurk-like paradigms further suggest that bilingual experience shapes audiovisual speech integration from early in development (Mercure, Bright, Quiroz, & Filippi, 2022). Investigating how multimodal information supports L2 comprehension in naturalistic settings is therefore essential to understanding L2 processing mechanisms.

Previous studies show that multimodal cues—such as facial expressions, gestures, and mouth movements—can modulate speech comprehension in both L1 and L2 speakers (Akker & Cutler, 2003; Drijvers, Vaitonytė, & Özyürek, 2019; Takahashi et al., 2018; Zhang, Frassinelli, Tuomainen, Skipper, & Vigliocco, 2021). However, most work has manipulated these cues in isolation, in ways that do not reflect how they co-occur in naturalistic contexts. This limitation in ecological validity raises the question of whether L2 speakers can benefit from co-occurring multimodal cues during real-world speech comprehension. To address this, we conducted an EEG experiment examining the effects of visual context on speech processing using naturalistic materials.

Current Study

In this study, we aimed to explore whether visual context makes speech processing more challenging by increasing cognitive demands for integration, or whether it facilitates speech processing by providing additional contextual information. Further, we wanted to explore whether this effect was modulated by language experience (i.e. monolingual vs. bilinguals), given that bilinguals typically have less experience than monolinguals with their shared language.

Two dimensions – acoustic and semantic information – of the stimulus were selected to measure how well each group tracks the information in continuous speech. Sensitivity to acoustic information was operationalized as tracking of the amplitude of the speech envelope, known alternately as *envelope tracking* or *audio tracking*. Envelope tracking is thought to capture the process by which ongoing cortical oscillations synchronize with slow amplitude fluctuations in speech, and is presumed to represent the initial neural encoding of speech (Giraud & Poeppel, 2012; Gross et al., 2013). Semantic information tracking, on the other hand, was operationalized as (lexical) surprisal tracking. Surprisal is the log-transformed conditional probability of a word based on the preceding context. It has previously been argued to index semantic processing as its neural correlates include the N400 ERP component

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

(Frank, Otten, Galli, & Vigliocco, 2015). Together, these two measures allow us to probe participants' sensitivity to both higher (i.e. semantic processing) and lower (i.e. envelope tracking) level aspects of the speech signal.

The temporal response function (TRF) method was then employed to estimate each individual's neural tracking performance. A temporal response function is a linear stimulus-response model that maps continuous stimuli such as speech to aspects of M/EEG signals. It has been shown that EEG responses can be predicted from the statistical measures of either acoustic or semantic features of the eliciting speech (forward modeling) (Lalor, Power, Reilly, & Foxe, 2009), or vice versa where properties of the speech can be predicted from EEG responses (backward modeling) (Lalor & Foxe, 2010). In this study, backward modeling was used to assess how accurately the selected feature could be reconstructed from neural activity. Better reconstruction accuracy was interpreted as better tracking of the features.

In summary, we collected scalp-recorded electroencephalogram (EEG) as monolingual and bilingual English speakers watched excerpts from Ted Talks in both audio-only and audio-visual formats. Then, the temporal response function (TRF) was employed to estimate how well each group tracked the audio envelope and surprisal in each condition (Crosse, Di Liberto, Bednar, & Lalor, 2016). Within each group, we asked 1) how does visual context affect envelope tracking and 2) how does visual context affect surprisal tracking by comparing the corresponding reconstruction accuracy in each condition. Significant differences in reconstruction accuracy as a function of stimulus format would suggest that envelope/surprisal tracking is modulated by visual context in the group. Alternatively, the absence of said effects may indicate that the group is less sensitive to or less reliant on visual context for envelope/surprisal tracking.

Materials and Methods

Participants

48 UCSD undergraduates (24 monolingual speakers, and 24 Mandarin-English bilingual speakers) were recruited to complete the study in exchange for extra credit in their cognitive science, linguistics, or psychology courses. The monolingual speakers all identified as native English speakers with no significant exposure to other languages, and the bilingual speakers identified as Mandarin/English bilingual speakers without significant exposure to any other languages. Bilingual speakers were accepted regardless of the order of acquisition of their two languages or the relative dominance between them. All participants were right handed, reported normal or corrected to normal vision and no history of reading or hearing difficulties. Informed consent was obtained from all participants prior to participating in the study.

Stimuli Design

5 video excerpts were taken from a corpus of Ted Talks in English. Videos were chosen based on the following criteria: 1) the talk was recorded post-2005 for clear video quality; 2) a single speaker talks for the entire duration of each presentation; 3) the speaker makes minimal use of videos/slide presentations to convey information; and 4) the camera is mainly focused on the speaker. The duration of each talk ranges from 7.7 to 9.3 minutes (mean duration = 8.3 mins, std = 0.7 mins). To promote attention to the material, each of the talks was broken down into several shorter clips varying from 1.3–2.6 minutes timed to end on a natural pause, resulting in 23 clips in all. Two versions were created of each clip – Audio-Only (AO) and Audio-visual (AV) conditions. In the audio-only condition, the visual component of the clip was replaced with a black screen with a white cross in the center, while the audio-visual condition retained the video's original form. The clips were then distributed to 2 different video lists, designed such that the audio-only and audio-visual conditions were counterbalanced across lists, and that each subject would see only a single version of each clip. The audio-only and audio-visual conditions alternated sequentially within each Ted Talk.

Comprehension Questions

At the end of each of the 5 talks, participants were prompted to answer a set of 3 multiple-choice questions intended to test their comprehension of the materials. Each multiple-choice question had 4 options, with a single correct answer. In total, 15 questions were answered across the 5 Ted Talks. The comprehension score for each participant was computed by dividing the total number of questions by the number of questions for which the participant chose the correct answer.

Category Verbal Fluency Test

A verbal category fluency task was administered to monolingual subjects in English and to bilingual subjects in both English and Mandarin. In this task, subjects were asked to produce as many items as possible from a specific semantic category within a 30 second time limit (Luo, Luk, & Bialystok, 2010; Rohrer, Wixted, Salmon, & Butters, 1995). Four different semantic categories were presented for each language. Two of the four categories were animate, while the other two were inanimate. The order of the categories was randomized. At the beginning of the task, participants had the opportunity to practice producing tokens for one semantic category that was unrelated to the main task categories (i.e., tools). All participant responses were recorded for later transcription and coding. Participants' performance was quantified by taking the sum of tokens across all categories within each language.

Experiment

Experimental Procedure After informed consent was obtained, the experimenter administered the verbal fluency test (monolingual participants) or tests (bilingual participants). Next, the participant was guided to a dimly lit, sound-proof room for the EEG portion of the experiment. Prior to the official EEG experiment, a practice session was conducted to

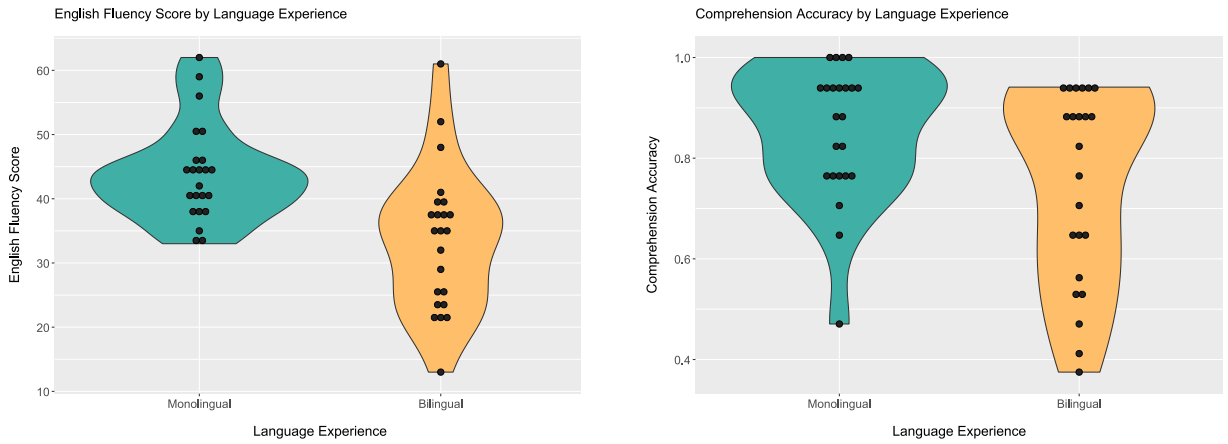

Figure 1: English fluency score and comprehension accuracy score distribution in monolingual and bilingual speakers.

demonstrate the format of the experiment, viz., sequential alternation between audio-only and audio-visual clips. The official EEG experiment comprised 5 blocks separated by self-paced breaks, each containing a complete Ted Talk excerpt. Participants were instructed to fixate on the white cross in the center of the monitor and listen to each clip in order to answer comprehension questions at the end of each video. In each block, participants viewed a sequence of clips that alternated between the audio-only (AO) and audio-visual (AV) excerpts; each clip was preceded by a white fixation cross in the center of the screen whose presentation duration varied randomly between 500 and 1000 ms. Following all the excerpts in a block, participants were prompted to answer the 3 comprehension questions about the Ted Talk they had just watched. Following each block, participants pressed a button to indicate when they were ready to proceed to the next Ted Talk.

EEG Acquisition and Preprocessing The EEG signals were recorded using 64 Ag/AgCl electrodes from the extended 10-20 standard system using a BioSemi ActiveTwo EEG system at a sampling rate of 2048Hz. The EEG signals were then band-pass filtered at [0.1Hz, 32Hz]. Artifact subspace reconstruction (ASR) and adaptive mixture independent component analysis (AMICA) were applied to remove artifacts (Chang, Hsu, Pion-Tonachini, & Jung, 2018; Hsu et al., 2018). Following artifact correction, the EEG data (resampled at 64Hz and 50Hz) was segmented into 23 epochs of varying length using timestamps that defined the onset and offset of each video clip. The 64Hz and 50Hz dataset was used for audio-envelope decoding and surprisal decoding, respectively. The resampling rate was decided based on the common denominator between EEG sampling rate and the stimulus sampling rate (e.g. original audio sampling rate: 44100 Hz; original surprisal sampling rate: 1000Hz). The preprocessing procedure was carried out using the EEGLAB toolbox on Matlab (Delorme & Makeig, 2004).

Temporal Response Function (TRF) Analysis

Stimulus Preprocessing To properly apply the TRF technique, the stimulus features must be extracted. The audio envelope was derived by transforming the audio source for each video clip into a univariate speech envelope, log-scaling the root-squared average amplitude across samples at 64Hz. This process was implemented by using the mTRFenvelope() function from the mTRF toolbox (Crosse et al., 2016).

The surprisal of every spoken word in the talks were extracted from the *davinci-002* model from Openai (T. Brown et al., 2020). Across all talks, The surprisal ranged from 1.89 to 6.24 (mean = 2.56), offering sufficient spread for meaningful analysis. All the surprisal measures were organized in a 1000Hz stream where the surprisal value was annotated with precise timing of each corresponding spoken word provided by Whisper (Radford et al., 2022) for each clip. The surprisals were annotated from the onset to the offset of the corresponding spoken word during the talk. Prior to the decoding analysis, the surprisals were downsampled to 50Hz to align with the EEG data.

Decoder Model Training The TRF technique was employed to model the relationship between EEG signals to audio envelope and surprisals (Crosse, Butler, & Lalor, 2015). In this case, the backward model (referred to below as the *decoder model*) was constructed to predict each stimulus feature from the EEG data. For every data point in the stimulus, a pre-selected window of the corresponding EEG signals (-100ms to 500ms for the audio envelope, 200ms to 700ms for surprisals) from all 64 channels was selected to construct a linear regression model to predict either the surprisal or the amplitude of the speech envelope at any given point in the au-

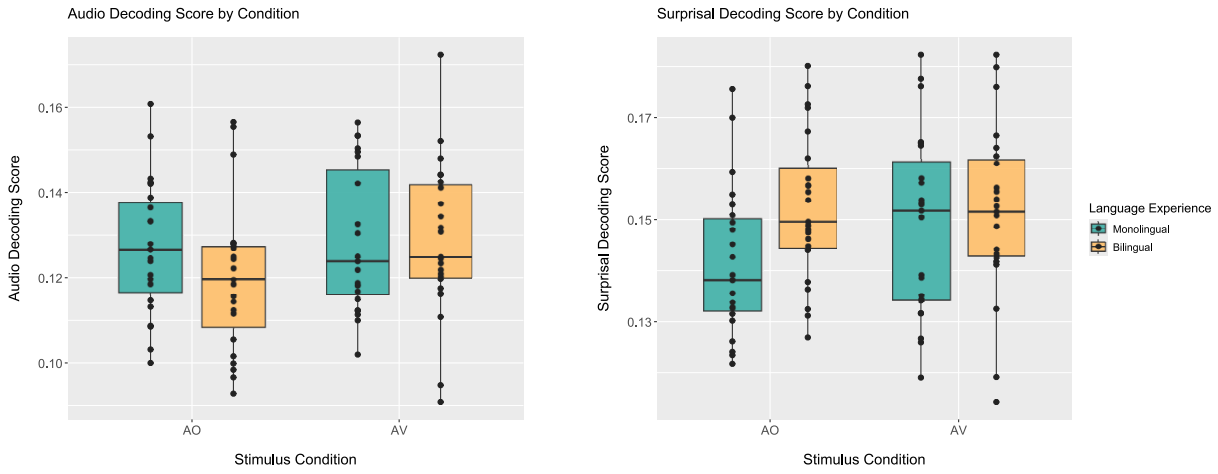

Figure 2: Audio Envelope (left) and surprisal (right) Decoding performance in two speaker groups across both conditions.

dio stream. In other words, the relationship between the stimulus feature and the EEG signal at any time point is described by a decoder model containing weighted EEG signals from the pre-selected time window. The time window was chosen to cover significant event related potentials (ERP) signatures of auditory processing (e.g. N1 and P2) and semantic processing (e.g. N400) (Fogarty, Barry, & Steiner, 2020; Kutas & Federmeier, 2011).

Prior to decoder model training, all stimuli and corresponding EEG signals were segmented into 10 second trials, resulting in a total of 235 trials. Separate decoder models were trained and constructed for each individual trial. To avoid data leakage, the target for each decoder model was first set aside as the testing trial so that it was never seen in the training phase. The first step of the decoder training session was to optimize the parameters of the decoder model in a leave-one-out cross validation fashion. Specifically, each of the 234 trials in the training set served as some point as a testing trial for one of the decoder models we constructed. The ridge parameters of the final decoder model were computed by averaging over all the decoder models assigned to the training trials. This model was then applied to the testing trial. Decoder accuracy was operationalized by the pearson's correlation score between the predicted stimulus signal (produced by the decoder) and the true stimulus signal for the testing trial. This correlation coefficient will be referred as decoding scores below. Separate training session were conducted for each individual subject and stimulus conditions (i.e. AO and AV conditions). This analysis was carried out using the multivariate temporal response function (mTRF) Matlab toolbox (Crosse et al., 2016).

Statistical Analysis The correlation coefficients from each testing trial in each participant were entered into a series of increasing complex Linear Mixed Effects (LME) models, examining the main effects of language experience and stimulus condition on the quality of decoder models derived from participants' neural activity. All models shared the same random effect structure, including random intercepts for the stimulus item (i.e. the stimulus clips) and subjects. The regression model with the lowest Akaike information criterion (AIC) was selected as the best model (Akaike, 1973). The audio-only condition (AO) and monolingual speakers were chosen as the baseline level for stimulus condition and language experience for all models, such that the coefficients for each predictor can be interpreted as the intercept-relative change as the result of having additional visual input or additional language experience (i.e. bilingual speakers).

Results

Behavioral Data

Categorical Verbal Fluency Test Performance on the verbal fluency test was determined by summing the total number of tokens produced across all four categories for each language. The result is shown in Figure 1. On average, the Welch's t-test revealed that the monolingual speakers produced more English tokens than the bilingual speakers (Monolingual: 44 ± 8 tokens, Bilingual: 34 ± 11 tokens; $t(38.71) = 3.72$, p-value < 0.001), indicating that monolingual speakers had higher English fluency than the bilingual speakers. Within the bilingual speakers, on average more English tokens were produced than Mandarin tokens (English: 34 ± 11 tokens, Mandarin: 27 ± 13 tokens). 5 out of 24 bilingual speakers produced more Mandarin tokens (i.e. Mandarin dominant), 7 produced more English tokens (i.e. English dominant), and the remaining 12 produced a similar number of tokens between both languages (i.e. balanced dominance).

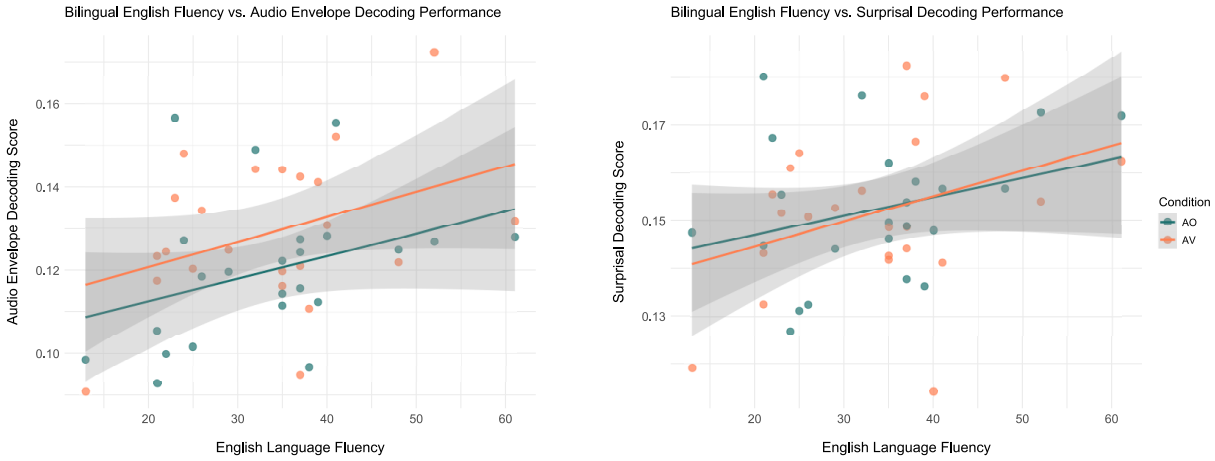

Figure 3: Correlation between English Fluency and decoding performance in bilingual speakers.

Table 1: Linear Mixed Effect Regression Model Result

Audio Envelope Decoding Decoding Accuracy∼ RV + LE*Condition				Surprisal Decoding Decoding Accuracy∼ RV + LE*Condition			
Predictors	Estimates	CI	p-value	Predictors	Estimates	CI	p-value
(Intercept)	12.70	[11.94, 13.46]	<**0.001**	*(Intercept)*	14.18	[13.44, 14.91]	<**0.001**
LE (B)	-0.73	[-1.66, 0.20]	0.125	LE (B)	1.03	[0.23, 1.83]	**0.011**
Con(AV)	0.20	[-0.30, 0.69]	0.439	Con(AV)	0.70	[0.11, 1.30]	**0.021**
LE(B) × Con(AV)	0.72	[0.01, 1.43]	**0.046**	LE(B) × Con(AV)	-0.65	[-1.50, 0.20]	0.132

RV: Random variables, LE:Language Experience, B: Bilingual, Con: Condition, AV: Audio-Visual

Comprehension Accuracy Comprehension accuracy is shown in Figure 1. All participants performed well above chance on the multiple choice questions, with an average accuracy of 80 % (SD: 17%). However, between the speaker groups, the Welch's t-test suggested that the monolinguals had higher accuracy rates than the bilinguals (Monolingual: 85 ± 13%, Bilingual: 75 ± 19%; t(38.71) = 2.22, p < 0.05).

Audio Decoding Performance

The best fitting model for audio decoding performance is the one including an interaction between language experience and stimulus condition (e.g. Audio Envelope Decoding Score ∼ Random Variables + Language Experience*Condition). The results are shown in Table1. This model reveals neither a main effect of AV condition nor that of bilingual language experience, but does indicate an interaction effect between stimulus condition and language experience (Estimates: 0.72, CI: [0.01, 1.43], $p = 0.046$). This model thus suggests that while monolingual speakers did not benefit from the visual input, audio envelope tracking in the bilinguals improved in the AV condition. Results can be seen in Figure 2.

Motivated by the interaction between bilingual speakers and stimulus condition, we wondered if the improvement from visual input was related to the large variance of language fluency in bilingual speakers. We constructed a follow-up LME model including both English fluency and Mandarin fluency (e.g. Audio Envelope Decoding Score ∼ Random Variables + English Fluency + Mandarin Fluency + Condition) to examine whether the improved performance in bilingual speakers can be explained away by different language fluency levels. The resultant model indicated that auditory decoding score was significantly related to stimulus condition (Estimates: 0.88, CI: [0.38, 1.38], $p = 0.001$), English fluency (Estimates: 0.06, CI: [0.00, 0.11], $p = 0.035$), but not Mandarin fluency (Estimates: -0.01, Confidence Interval (CI): [-0.06, 0.04], $p = 0.68$). Audio tracking increased as a function of English fluency, but not Mandarin fluency. Moreover, even when controlling for differences in English fluency, audio tracking in bilinguals was better with visual input. These effects can be seen in the left-hand panel of Figure 3.

Surprisal Decoding Performance

As for the audio decoding analysis, a series of increasingly complex LME models of surprisal decoding performance

were constructed. Statistical model comparison indicated that the best model of surprisal decoding scores was the one that included an interaction effect between language experience and condition (e.g. Surprisal Decoding Score$\sim$ Random Variables + Language Fluency * Condition). This model revealed a main effect of the the AV condition (Estimates: 0.70, CI: [0.11, 1.30], p = 0.021) indicating a beneficial effect of the visual cues on surprisal decoding in the monolingual reference group, and a main effect of bilingual experience (Estimates:1.03, CI: [0.23, 1.83], p = 0.011) indicating better surprisal decoding on the audio-only trials among the bilingual speakers. However, the interaction term failed to reach significance (Estimates:-0.65, CI: [-1.50, 0.20], p = 0.132), suggesting the bilinguals failed to benefit from the visual input. The surprisal decoding scores are plotted in Figure 2 and the output of the mixed effects model is presented in Table 1.

Similar to the auditory analysis, we again asked if surprisal tracking performance in bilingual participants was related to their language fluency level. Accordingly, we constructed an analogous follow-up model (e.g. Surprisal Decoding Score $\sim$ Random Variables + English Fluency + Mandarin Fluency + Condition). This post hoc model revealed that bilinguals' surprisal tracking improved with their English fluency (Estimates:0.05, CI: [0.01, 0.08], p = 0.009), but not their Mandarin fluency (Estimates:-0.01, CI: [-0.04, 0.02], p = 0.517). The null effect of AV condition in this model (Estimates:0.08, CI: [-0.53, 0.69], p = 0.794) confirms that visual input in the AV condition did not facilitate surprisal tracking among the bilingual speakers. The positive relationship between English fluency and surprisal decoding can be seen in the right-hand panel of Figure 3.

Discussion

In the current study, we explore how language experience influences neural tracking of the speech envelope and of the surprisal of words in the speech contingent on the presence or absence of concurrent visual input. To do this, we employed the TRF technique to decode EEG signals recorded from monolingual and bilingual adults listening to audio-only and audio-visual excerpts from Ted Talks. We then applied a series of linear mixed-effect regression models to analyze the relationship of tracking performance with the presence of visual cues in the stimuli and with the language experience of the participants. We found that visual input influenced different aspects of speech processing in our two groups of participants. Among bilinguals, visual context led to better tracking of the audio envelope, but conferred no apparent advantage for suprisal tracking. Among monolinguals, visual context led to better tracking of lexical surprisal, but conferred no apparent advantage for audio tracking.

Visual Cues in Audio Tracking versus Surprisal Tracking

As noted above, we find that visual cues facilitated low level auditory processing in bilinguals, but provided little impact on higher level semantic processing. This finding is consistent with prior studies that have reported that, relative to L1 listeners, L2 listeners are more sensitive to visual cues during speech processing (Drijvers & Özyürek, 2018; Birulés et al., 2020), but benefit less from mouth movements and meaningful gestures in tests of language comprehension (Drijvers et al., 2019; Drijvers & Özyürek, 2019). We suggest that one consequence of bilingual participants' more limited experience with English is that lower-level aspects of processing may demand more of their attention, thereby reducing the cognitive resources available for processing meaning. In line with this suggestion, Brown and Strand (2019) found that while visual information facilitates word recognition, participants' overall cognitive listening effort increased. If visual input boosted our bilinguals' audio decoding scores while simultaneously increasing cognitive load, the reduction in resources available for higher level processing might explain why surprisal tracking levels remained constant. This is also supported by the positive relationship between English fluency and decoding scores for both the sound envelope and lexical surprisal.

Surprisal Tracking in Monolinguals versus Bilinguals

Another curious result of the present study was that suprisal decoding scores during audiovisual clips were similar in both groups, and were actually better in the bilinguals during the audio-only clips. One potential explanation may lie in the construct validity of this measure. That is, perhaps the surprisal measures we obtained from *davinci-002* do not provide an accurate estimation of the surprisal values for words in our multi-modal corpus due to the prevalence of textual input in GPT-3's training set. Indeed, it may be that the surprisal measures from the language model are more similar to those of our bilingual speakers who may also have more experience with written English. Alternatively, if we accept that the surprisal measures used here do provide a reasonable estimate of the unpredictability of words in our corpus, group differences observed in the present study may result because bilingual participants make greater use of predictive processing mechanisms, and therefore are more sensitive to surprisal measures than their monolingual peers. More research is needed to clarify the factors that influence surprisal tracking in multi-modal discourse comprehension.

In conclusion, our study is one of the first to investigate multi-modal speech processing by directly analyzing the relationship between neural signals and naturalistic materials. Our results showed that visual information facilitates different aspects of speech processing in monolingual and bilingual speakers. However, in view of the limitations of the paradigm used here (e.g., the coarse manipulation of visual context, the use of only two dependent measures, etc.), more research is needed to characterize how language experience impacts the perception and comprehension of speech in multi-modal contexts.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Proceedings of the 2nd international symposium on information theory* (pp. 267–281). Akadémiai Kiadó. doi: 10.1007/978-1-4612-1694-0$_1$5

Akker, E., & Cutler, A. (2003). Prosodic cues to semantic structure in native and nonnative listening. *Bilingualism: Language and Cognition*, *6*, 81–96. doi: 10.1017/S1366728903001067

Birulés, J., Bosch, L., Pons, F., & Lewkowicz, D. J. (2020). Highly proficient l2 speakers still need to attend to a talker's mouth when processing l2 speech. *Language, Cognition and Neuroscience*, *35*(10), 1314–1325. doi: 10.1080/23273798.2020.1762905

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Brown, V. A., & Strand, J. F. (2019). About face: Seeing the talker improves spoken word recognition but increases listening effort. *Journal of Cognition*, *2*(1), 44. doi: 10.5334/joc.89

Chang, C.-Y., Hsu, S.-H., Pion-Tonachini, L., & Jung, T.-P. (2018). Evaluation of artifact subspace reconstruction for automatic eeg artifact removal. In *2018 40th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 1242–1245). IEEE. doi: 10.1109/EMBC.2018.8512543

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(42), 14195–14204. doi: 10.1523/JNEUROSCI.1829-15.2015

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*, 604. doi: 10.3389/fnhum.2016.00604

Dahl, T. I., & Ludvigsen, S. (2014). How i see what you're saying: The role of gestures in native and foreign language listening comprehension. *Modern Language Journal*, *98*, 813–833. doi: 10.1111/modl.12124

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Drijvers, L., Vaitonytė, J., & Özyürek, A. (2019). Degree of language experience modulates visual attention to visible speech and iconic gestures during clear and degraded speech comprehension. *Cognitive Science*, *43*(10), e12789. doi: 10.1111/cogs.12789

Drijvers, L., & Özyürek, A. (2018). Native language status of the listener modulates the neural integration of speech and iconic gestures in clear and adverse listening conditions. *Brain and Language*, *177–178*, 7–17. doi: 10.1016/j.bandl.2018.01.004

Drijvers, L., & Özyürek, A. (2019). Non-native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension. *Language and Speech*, *63*(2), 209–220. Retrieved from https://pubmed.ncbi.nlm.nih.gov/30795715/ doi: 10.1177/0023830919831311

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2023). *Ethnologue: Languages of the world* (26th ed.). Dallas, Texas: SIL International. Retrieved from https://www.ethnologue.com/

Fogarty, J. S., Barry, R. J., & Steiner, G. Z. (2020). Auditory stimulus- and response-locked erp components and behavior. *Psychophysiology*, *57*(5), e13538. doi: 10.1111/psyp.13538

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The erp response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1-11. doi: https://doi.org/10.1016/j.bandl.2014.10.006

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. doi: 10.1038/nn.3063

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biology*, *11*(12), e1001752. doi: 10.1371/journal.pbio.1001752

Hsu, S.-H., Pion-Tonachini, L., Palmer, J., Miyakoshi, M., Makeig, S., & Jung, T.-P. (2018). Modeling brain dynamic state changes with adaptive mixture independent component analysis. *NeuroImage*, *183*, 47–61. doi: 10.1016/j.neuroimage.2018.07.019

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual Review of Psychology*, *62*, 621–647. doi: 10.1146/annurev.psych.093008.131123

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *The European Journal of Neuroscience*, *31*(1), 189–193. doi: 10.1111/j.1460-9568.2009.07055.x

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. *Journal of Neurophysiology*, *102*(1), 349–359. doi: 10.1152/jn.90896.2008

Lee, A., Perdomo, M., & Kaan, E. (2019). Native and second-language processing of contrastive pitch accent: An erp study. *Second Language Research*, *36*(4), 503–527. doi: 10.1177/0267658319838300

Lin, Y.-L. (2021). Gestures as scaffolding for l2 narrative recall: The role of gesture type, task complexity, and working memory. *Language Teaching Research*. doi: 10.1177/13621688211044584

Luo, L., Luk, G., & Bialystok, E. (2010). Effect of language proficiency and executive control on verbal fluency performance in bilinguals. *Cognition*, *114*(1), 29–41.

Mercure, E., Bright, P., Quiroz, I., & Filippi, R. (2022). Effect of infant bilingualism on audiovisual integration in a mcgurk task. *Journal of Experimental Child Psychology*, *217*, 105351. doi: 10.1016/j.jecp.2021.105351

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision.* Retrieved from https://arxiv.org/abs/2212.04356

Rohrer, D., Wixted, J. T., Salmon, D. P., & Butters, N. (1995). Retrieval from semantic memory and its implications for alzheimer's disease. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1127–1139.

Rosenberg, A., Hirschberg, J. B., & Manis, K. (2010). Perception of english prominence by native mandarin chinese speakers. In *Speech prosody 2010.* Retrieved from https://doi.org/10.7916/D8BR91N2

Takahashi, C., Kao, S., Baek, H., Yeung, A. H., Hwang, J., & Broselow, E. (2018). Native and non-native speaker processing and production of contrastive focus prosody. *Proceedings of the Linguistic Society of America*, *3*(1), 35:1–13. doi: 10.3765/plsa.v3i1.4340

Zhang, Y., Frassinelli, D., Tuomainen, J., Skipper, J. I., & Vigliocco, G. (2021). More than words: word predictability, prosody, gesture and mouth movements in natural language comprehension. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1955), 20210500. doi: 10.1098/rspb.2021.0500