*INTERSPEECH 2019*
September 15–19, 2019, Graz, Austria

# Effects of Natural Variability in Cross-Modal Temporal Correlations on Audiovisual Speech Recognition Benefit

*Kaylah Lalonde*

Boys Town National Research Hospital, Omaha, Nebraska, USA

Kaylah.lalonde@boystown.org

## Abstract

In audiovisual (AV) speech, correlations over time between visible mouth movements and the amplitude envelope of auditory speech help to reduce uncertainty as to when peaks in the auditory signal will occur. Previous studies demonstrated greater AV benefit to speech detection in noise for sentences with higher cross-modal correlations than sentences with lower cross-modal correlations.

This study examined whether the mechanisms that underlie AV detection benefits have downstream effects on speech recognition in noise. Participants were presented 72 sentences in noise, in auditory-only and AV conditions, at either their 50% auditory speech recognition threshold in noise (SRT-50) or at a signal-to-noise ratio (SNR) 6 dB poorer than their SRT-50. They were asked to repeat each sentence. Mean AV benefit across subjects was calculated for each sentence. Pearson correlations and mixed modeling were used to examined whether variability in AV benefit across sentences was related to natural variation in the degree of cross-modal correlation across sentences.

In the more difficult listening condition, higher cross-modal correlations were associated with higher AV sentence recognition benefit. The relationship was strongest in the 0.8-2.2 kHz and 0.8-6 kHz frequency regions. These results demonstrate that cross-modal correlations contribute to variability in AV speech recognition in noise.

**Index Terms**: speech recognition, multimodal, audiovisual

## 1. Introduction

AV speech perception is more accurate and less effortful than auditory-only speech perception, especially when the auditory signal is degraded (i.e., by acoustically noisy backgrounds) [1]. The increased accuracy and efficiency of speech processing in the presence of visual speech is called AV speech enhancement.

Auditory and visual speech signals are correlated in multiple ways [2, 3]. Thus, there are multiple mechanisms by which visual speech enhances speech recognition [4-6]. One potential mechanism is that correlations over time between visible mouth movements and acoustic energy (where an open mouth corresponds to relatively greater acoustic amplitudes) help to direct auditory analysis [4-6]. More specifically, the cross-sectional area of the front portion of the vocal tract is proportional to acoustic intensity, especially in the spectral region of the second and third formants (F2 and F3) [7-9]. Correlations between mouth movements and the acoustic envelope help to reduce uncertainty as to when onsets and peaks in the auditory signal will occur [4,10].

Grant and Seitz [4] demonstrated the benefit of cross-modal correlations by measuring auditory-only and AV detection of spoken sentences. Participants were presented two observation intervals containing white noise and asked to judged which of the two intervals contained auditory speech. On auditory-only trials, no visual signal was presented; On AV trials, the same visual speech signal was presented in both observation intervals. Adults were able to detect auditory sentences at poorer SNRs when presented the congruent visual speech than when presented mismatched visual speech or no visual speech. The AV detection benefit results from the use of visual temporal cues, not from knowing the content of the speech signal: Providing orthographic information—rather than visual speech—did not provide the same benefits [4], and AV detection benefits are observed for speech spoken in a foreign language [11].

Importantly, the AV detection benefit in Grant and Seitz's study [4] varied across the 3 sentences that served as target stimuli. Subsequent analysis of the AV stimuli indicated a higher cross-modal correlation between the acoustic amplitude envelope and the area of the mouth opening for the high benefit sentence than the low benefit sentence, especially in the F2 spectral region (0.8-2.2 kHz). In a subsequent study, Grant [12] confirmed the importance of the correlation between the F2 region and the mouth movements by repeating the detection experiment with band-pass filtered auditory speech. Significant AV detection benefit was observed for unfiltered speech and speech from the F2 spectral region, but not for speech from the F1 (0.1-0.8 kHz) and F3 (2.2-6.0 kHz) spectral regions. In summary, participants used correlations between the area of the mouth opening and acoustic energy in the F2 spectral region to enhance speech detection.

The mechanisms underlying AV detection benefit are believed to enhance speech recognition as well [13, 14]. Such a belief is supported by the hierarchical nature of speech perception. One must detect a signal (be aware of its presence) in order to discriminate its component features and activate the appropriate lexical representations in long-term memory. Thus, if cross-modal correlations help detect auditory speech in noise, they should effectively improve SNRs, making phonetic features more discriminable and lexical representations easier to access. However, detection benefits are improvements in threshold-level processing. Therefore, the effect of cross-modal correlations may not always be evident when testing speech recognition at supra-threshold levels. Further, the benefit of cross-modal correlations is likely more difficult to observe in speech recognition, due to the added influence of higher-level processes (i.e., use of phonetic and lexical knowledge [15, 16] and sentence context [17, 18]).

Few studies have directly assessed whether the mechanisms underlying AV speech detection benefit generalize to speech identification and recognition. One previous study demonstrated effects of visual temporal cues on speech sound discrimination/identification [13]. An



http://dx.doi.org/10.21437/Interspeech.2019-2931

ambiguous visual speech signal helped French-speaking adults to detect an auditory pre-voicing cue, resulting in better discrimination between syllables with initial voiced plosives (/dy/, /du/, /gy/, /gu/) and other syllables with either voiceless plosives (/ty/, /tu/, /ky/, /ku/) or no consonant (/y/, /u/). The advantage was observed even when there was no phonetic information in the visual signal (when the same visual speech signal was paired with all auditory consonants).

The purpose of this study was to examine whether the cross-modal correlation mechanism that underlies AV sentence detection benefit also applies to sentence recognition in noise. To that end, experiments were carried out to 1) characterize the natural variability in cross-modal correlations across a set of 72 sentences, and 2) relate variability in cross-modal correlations to variability in AV benefit to speech recognition across sentences. To minimize effects of higher-level processes—and improve the potential sensitivity of the sentence recognition task to cross-modal correlations—syntactically correct but semantically meaningless sentences were used. Additionally, testing was completed at two levels of auditory performance. One group of participants was tested at relatively easy SNRs typical of AV speech recognition studies (at subjects' 50% auditory speech recognition threshold in noise [SRT-50]) [19]. Another group was tested at relatively hard SNRs, 6 dB poorer than participants' SRT-50 (and, thus closer to the level at which effects of cross-modal correlation are observed in AV detection experiments). We expected cross-modal correlations (particularly in the high-frequency regions corresponding to F2) to predict stimulus-specific differences in AV enhancement. Additionally, we expected a stronger relationship for sentences presented closer to threshold (harder SNRs) than for sentences presented at supra-threshold levels (easier SNRs).

## 2. Method

### 2.1. Stimuli and Stimulus Analysis

The stimulus set included professional AV recordings of 72 syntactically correct but semantically anomalous sentences [20] from 4 talkers (10 to 24 sentences per talker). E.g., "The twirl hides the easy laundry."

Final Cut Pro [21] was used to save each sentence as a separate video file. FaceScanner software [22] was used to measure the position of 8 points at the perimeter of the opening of the talker's mouth, in each frame of the video files. FaceScanner data were imported to MATLAB, where the polyarea function was used to calculate the area of the mouth opening in each frame [23]. An example of the output of this analysis for a single sentence is shown in black in Figure 1.

Adobe Audition [24] was used to manually determine the start and end of the acoustic sentences and to equate the total RMS amplitude of each sentence. Equalized audio signals were imported into MATLAB [23] and filtered into 5 frequency regions using the bandpass function in the Signal Processing Toolbox [25]. The filter conditions included 1) unfiltered, 2) F1 region, 0.1 to 0.8 kHz, 3) F2 region, 0.8 to 2.2 kHz, 4) F3 region, 2.2 to 6 kHz, and 5) F2+F3 region, 0.8 to 6 kHz. For each filter condition, we calculated the RMS of the 33.367 ms portion of the auditory signal corresponding to each video frame. An example of the output of this analysis for a single sentence is shown by the colored lines in Figure 1.

For each filter condition, a 10-frame (333.67 ms) moving window was used to calculated the instantaneous correlation between the area of the mouth opening in each frame and the RMS of the corresponding portion of the auditory signal. The mean instantaneous correlation over the duration of the acoustic sentence was used as a metric of the strength of the cross-modal correlation. To account for the fact that listeners can tolerate some AV asynchrony in speech perception, cross-modal correlations were calculated at +/-5 frames of visual lead [26, 27]. Overall, correlations were best with 1 frame of visual lead (as in [4]), but there was some variation. For the current study, the highest correlation between 0, 1, and 2 frames of visual lead was chosen for each sentence. Figure 2 shows the timewise instantaneous correlation between the area of the mouth opening and the auditory amplitude envelope for each filter condition, with 1 frame of visual lead, for a single sentence. Descriptive statistics regarding the mean instantaneous cross-modal correlation over the duration of sentences for each filter condition are provided in Table 1.

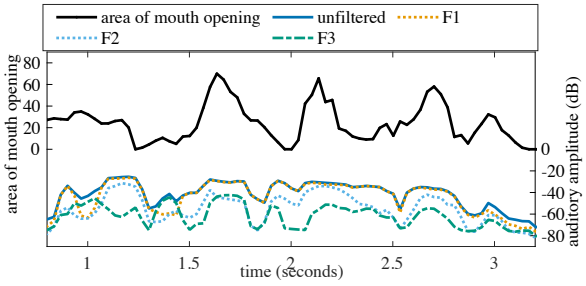


Figure 1: *Example of data extracted for stimulus analysis. Area of the mouth opening and auditory amplitude envelopes in the unfiltered condition, and F1, F2, and F3 spectral regions for one sentence.*

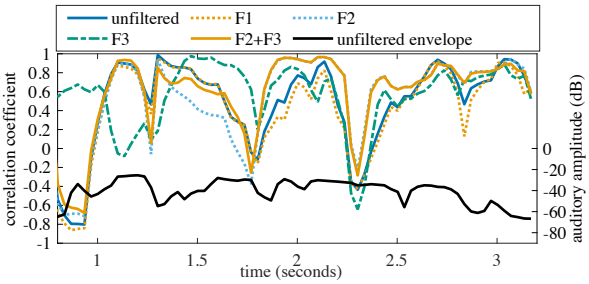


Figure 2: *Instantaneous cross-modal correlations for each filter condition are shown in color. The auditory amplitude envelope of the unfiltered signal is in black.*

Table 1: *Descriptive statistics regarding the cross-modal correlations for each filter condition.*

| Condition | Mean (S.D.) Correlation | Min – Max Correlation |
|---|---|---|
| Unfiltered | 0.342 (0.146) | 0.021 – 0.604 |
| F1 (0.1 – 0.8 kHz) | 0.340 (0.146) | 0.038 – 0.621 |
| F2 (0.8 – 2.2 kHz) | 0.340 (0.152) | 0.060 – 0.691 |
| F3 (2.2 – 6 kHz) | 0.395 (0.131) | 0.043 – 0.700 |
| F2+F3 (0.8 – 6 kHz) | 0.373 (0.139) | 0.095 – 0.671 |

### 2.2. Behavioral Experiments

Behavioral experiments were used to determine whether natural variation in the cross-modal correlation between visible mouth movements and the auditory amplitude envelope

of speech would predict stimulus differences in the degree to which visual speech improves sentence recognition. Participants completed a sentence recognition task in auditory-only and AV conditions, at individually-set SNRs.

### 2.3. Participants

Thirty-four adult native English speakers (18 to 35 years) participated in the study. All had normal hearing thresholds (screened at 20 dB HL bilaterally from 0.25 to 8 kHz), and normal or corrected-to-normal visual acuity.

### 2.4. Procedures

Research was approved by the Institutional Review Board at Boys Town National Research Hospital. Participants were tested in a large double-walled sound booth. Specialized software in Max 7 [28] on a Mac Pro computer was used to present stimuli, control the experiment, and collect responses. Stimuli routed to a 68.6-cm high-definition monitor and to loudspeakers positioned 45 degrees left and right of the monitor. Faces presented on the screen extended approximately 19 cm in height and 15.3 cm in width.

A modified speech reception threshold procedure was used to estimate SRT-50 for each subject [29]. Low-predictability sentences from the Speech Perception in Noise test [30] were presented auditorily at 65 dB SPL. Noise level varied adaptively (1-down/1-up) from trial to trial [31]. Step size decreased from 18 dB to 3 dB, and testing ended after 8 reversals. SRT-50 was the average SNR at the last 5 reversals.

Once a participant-specific SNR was established, participants were presented each of the 72 unfiltered sentences in a random order in speech-spectrum noise, in auditory and AV conditions, with order of test modality counterbalanced. Each sentence was presented once in each modality. Sixteen participants were tested at the easy SNR corresponding to their SRT-50, with target stimuli at 62 dB SPL. Eighteen participants were tested at the hard SNR, 6 dB below their SRT-50, with target stimuli at 56 dB SPL. In auditory-only conditions, no visual stimulus was presented. In AV conditions, a congruent video of the talker was presented. Participants repeated each sentence to the best of their ability, and the experimenter typed their response. Sentences were scored for number of keywords correctly recognized.

## Results

### 2.5. Behavioral Outcomes

Table 2 summarizes the auditory and AV accuracy outcomes for the easy and hard SNRs. A 2 x 2 mixed ANOVA was used to examine effects of modality (auditory and AV) and SNR (easy, hard) on sentence recognition accuracy. Participants in the easy SNR condition were tested at higher SNRs (Mean = -2.7 dB, S.D. = 1.7 dB) than participants tested in the hard SNR condition (Mean = -8.9 dB, S.D. = 1.5 dB). Consequently, participants tested at the easy SNR recognized keywords from the sentences more accurately overall than those tested at the hard SNRs ($F_{1,63}$ = 33.52, $p < 0.001$). Participants also recognized more keywords in the AV condition than in the auditory-only condition ($F_{1,63}$ = 26.54, $p < 0.001$). AV benefit did not differ significantly across groups, but mean AV benefit was greater for the hard SNR than for the easy SNR.

Table 2: *Mean (and standard deviation) keyword recognition accuracy across participants*

| Condition | Accuracy |
|---|---|
| Easy SNR, Auditory | 66.7% (12.6%) |
| Easy SNR, AV | 80.6% (9.4%) |
| Easy SNR, Benefit | 13.9% (9.2%) |
| Hard SNR, Auditory | 22.3% (15.5%) |
| Hard SNR, AV | 46.3% (20.6%) |
| Hard SNR, Benefit | 24.0% (11.9%) |

### 2.6. Relating AV benefit to cross-modal correlations

We calculated AV benefit for each sentence at each level of difficulty (easy SNR, hard SNR). AV benefit was defined as the mean difference between auditory-only and AV keyword recognition accuracy across participants. In the easy SNR, 13 AV sentences were recognized with 100% accuracy. Data for these sentences were excluded from the easy SNR analysis, because benefit may have been limited by ceiling performance. Pearson product moment correlations were used to examine the relationship between AV benefit and cross-modal correlation for each filter condition and SNR difficulty.

Correlation coefficients are shown in Table 3. Significant correlations were observed for the harder SNR in the unfiltered condition, r = 0.26, p = 0.029, F2 spectral region, r = 0.30, p = 0.012, and F2+F3 region, r = 0.37, p =0.002. Results are shown in Figure 3. No significant correlations were observed for the easier SNR. In addition, neither auditory-only nor AV accuracy significantly correlated with the cross-modal correspondences.

Table 3: *Correlations across sentences between AV benefit and Mean Instantaneous Cross-Modal Correlation, * $p < 0.05$, ** $p < 0.01$*

| Filter condition | Easy SNR | Hard SNR |
|---|---|---|
| Unfiltered | 0.13 | 0.26* |
| F1 (0.1 – 0.8 kHz) | 0.10 | 0.19 |
| F2 (0.8 – 2.2 kHz) | -0.03 | 0.30* |
| F3 (2.2 – 6 kHz) | 0.07 | 0.16 |
| F2+F3 (0.8 – 6 kHz) | 0.00 | 0.37** |

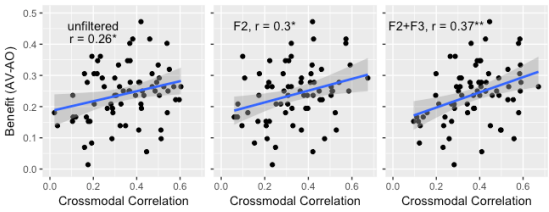


Figure 3: *Scatter plots show AV benefit for the hard SNR as a function of cross-modal correlation calculated from the unfiltered signal, F2 region, and F2+F3 region. Each marker represents a sentence.*

Participants encountered each stimulus twice, once in each modality. Linear mixed modeling was used to examine whether test order affected the correlations in Table 3 and Figure 3. Data from each SNR difficulty and filter condition were analyzed separately using the *lmer* function in the lme4 package in R [32, 33]. Mixed linear models with a random intercept for sentence were used to examine the effects of test order (coded categorically as auditory first, AV first) and

mean instantaneous cross-modal correlation on AV benefit. Results are shown in Figure 4.

In the easy SNR condition, final models included a significant effect of test order (Beta = -0.115, t = -7.951, $p < 0.001$, df = 73). There was greater AV benefit in group tested in the auditory-only condition first, reflecting stimulus familiarity and/or practice effects. For the easy SNR, there was no effect of cross-modal correlation and no interaction between test order and cross-modal correlation for any filter condition.

In the difficult SNR condition, the final model for all filter conditions included an effect of test order (Beta = -0.106, t = 6.752, $p \leq 0.001$, df = 72). The final models for the unfiltered condition, F2 spectral region, and F2+F3 spectral region also included a significant effect of cross-modal correlation (Beta $\geq$ 0.160, $t \geq 2.259$, $p \leq 0.027$, df = 72). Greater benefit was observed for stimuli with greater cross-modal correlations. No significant interactions were observed.

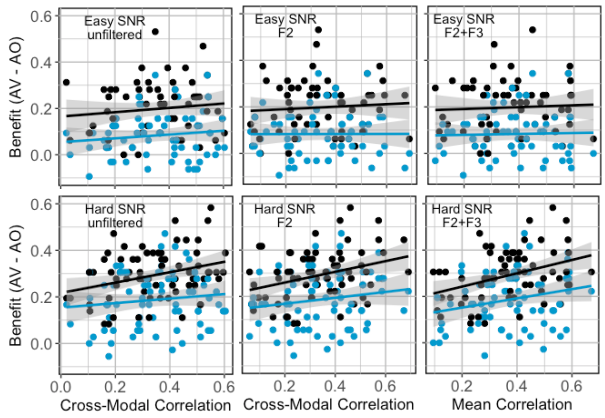


Figure 4: *AV benefit as a function of cross-modal correlation calculated from the unfiltered signal, F2 region, and F2+F3 region. Each marker represents a sentence. Color represents test-order, where black indicates auditory first and blue indicates AV first.*

## 3. Discussion

The results of this study demonstrate that the mechanisms underlying AV detection benefits also enhance speech recognition in noise. Specifically, cross-modal correlations over time between visible mouth movements and acoustic speech energy—especially in the F2 spectral region—contribute to AV enhancement of speech recognition. These results are consistent with neurophysiological evidence indicating that cross-modal correlations help the auditory cortex to track the amplitude envelope of speech [34-37]. Results are also consistent with Berthommier's phonetically neutral model of audio-visual interactions, which suggests that low-level AV enhancements are based on the product between the auditory and visual-predicted amplitude envelopes of speech signals [14]. Consistent with the current study and previous AV detection experiments [4, 10, 11], such tracking would theoretically improve both threshold-level detection and supra-threshold recognition of speech in noise.

The effects of cross-modal correlation were only observed at relatively difficult SNRs, not at SNRs typical of AV speech recognition studies. It is possible that listeners employ this mechanism only when listening to speech under highly degraded acoustic conditions. However, it is also possible that there were more opportunities for individual subjects to reach ceiling-level performance in AV conditions at the easy SNR. Ceiling performance would limit the maximum possible AV benefit, potentially masking the effects of cross-modal correlations at easier SNRs.

Few studies have examined the relationship between AV enhancement of speech recognition and natural variation in the physical properties of AV speech stimuli. One previous study examined correlations between visual entropy—a measure of visual speech information based on the running sum of color differences between successive video frames—and mid-frequency (1.5-2.5 kHz) speech energy across 60 sentences [38]. Kim and colleagues [38] related the cross-modal correlations to AV enhancement of sentence recognition, tested at -3 dB SNR. Consistent with the easy SNR condition in the current study, they observed a weak relationship between the cross-modal measure and AV benefit. In addition to testing at a relatively easy SNR, Kim et al. [38] used meaningful sentences that provide contextual information. This may have masked some effects of the stimulus characteristics. Future studies examining the relationship between physical characteristics of AV speech and behavioral measures of AV speech recognition benefit may profit from the use of low context stimuli and from testing at multiple levels of listening difficulty.

This study demonstrated the importance of cross-modal correlation between visible mouth movements and auditory amplitude envelopes for AV enhancement of speech recognition. Future studies will examine what happens to AV speech enhancement when these cross-modal correlations are disrupted. Wide dynamic range compression (WDRC) is a key feature in hearing aids and a primary means of improving auditory speech intelligibility in individuals with sensorineural hearing loss [39, 40]. However, WDRC also distorts the amplitude envelope of auditory speech [41]. Future studies will examine the degree to which WDRC-induced distortion of the amplitude envelope of auditory speech disrupts cross-modal correlations and—consequently—affects listeners' ability to use cross-modal correlations to enhance AV speech detection and recognition.

## 4. Conclusions

This study demonstrates a relationship between physical characteristics of AV speech and behavioral measures of AV speech recognition benefit. AV enhancement of speech recognition depends on natural variation in the degree of cross-modal correlation between visible mouth movements and the amplitude envelope of auditory speech, particularly in the F2 spectral region. Results suggest that visible mouth movements help to track the amplitude envelope of auditory speech in the F2 spectral region. Relatively difficult listening conditions are required to observe this relationship.

## 5. Acknowledgements

This research was funded by an IDeA CTR pilot grant (NIH-NIGMS, 1 U54 GM115458) and supported by the Technical Core within Boys Town National Research Hospital (NIH-NIGMS P20 GM109023). Seth Bashford, Tim Vallier, and Denis Fitzpatrick contributed to hardware and software development. Jamie Petersen and Nancy He collected behavioral data.

# 6. References


[1] W.|H.|Sumby and I.|Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[2] H.|Yehia, P.|Rubin, and E.|Vatikiotis-Bateson, "Quantitative association of vocal tract and facial behavior," *Speech Communication,* vol. 26, pp. 23-43, 1998.

[3] K.|G.|Munhall, and E.|Vatikiotis-Bateson, "Spatial and temporal constraints on audiovisual speech perception," *The Handbook of Multisensory Processes,* Cambridge, MA: The MIT Press, 2004.

[4] K.|W.|Grant and P.|Seitz, "The use of visible speech cues for improving auditory detection of spoken sentences," *Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1197–1208, 2000.

[5] J.|Peelle and M.|S.|Sommers, "Prediction and constraint in audiovisual speech perception," *Cortex,* vol. 68, pp. 169-181, 2005.

[6] Q.|Summerfield. "Some preliminaries to a comprehensive account of audio-visual speech perception," *Hearing by Eye: The Psychology of Lipreading,* Hillsdale, NJ: Lawrence Erlbaum, 1987.

[7] G.|Fairbanks. "A physiologic correlative of vowel intensity," *Communications Monographs,* vol. 17, no. 4, pp. 390-395, 1950.

[8] K.|N.|Stevens and A.|S.|House. "Development of a quantitative description of vowel articulation," *Journal of the Acoustical Society of America,* vol., 27, no. 3, pp. 484-493, 1955.

[9] K.|N.|Stevens and A.|S.|House. "An acoustical theory of vowel production and some of its implications," *Journal of Speech and Hearing Research,* vol. 4, no. 4, pp 302-320, 1961.

[10] J.|Kim and C.|Davis. "Investigating the audio-visual speech detection advantage," *Speech Communication,* vol. 44, no. 1-4, pp. 19-30, 2004.

[11] J.|Kim and C.|Davis. "Hearing foreign voices: does knowing what is said affect visual-masked-speech detection?" *Perception,* vol. 32, no. 1, pp. 111-120, 2003.

[12] K.|W.|Grant. "The use of visible speech cues for improving auditory detection of spoken sentences," *Journal of the Acoustical Society of America,* vol., 108, no. 3, pp. 1197-1208, 2000.

[13] J.|L.|Schwartz and F.|Berthommier. "Seeing to hear better: evidence for early audio-visual interactions in speech identification," *Cognition,* vol. 93, no. 2, pp. 69-78, 2004.

[14] F. |Berthommier. "A phonetically neutral model of the low-level audio-visual interaction," *Speech Communication,* vol. 44, no. 1-4, pp. 31-41, 2004.

[15] P.|A.|Luce and D.|B.|Pisoni. "Recognizing spoken words: the neighborhood activation model," *Ear and Hearing,* vol. 19, no. 1, pp. 1-36, 1998.

[16] N.|Tye-Murray, M.|S.|Sommers, and B.|Spehar. "Auditory and visual lexical neighborhoods in audiovisual speech perception," *Trends in Amplification,* vol. 11, no. 4, pp. 233-241, 2007.

[17] R.|W.|Hutcherson, D.|D.|Dirks, and D.|E.|Morgan. "Evaluation of the speech perception in noise (SPIN) test," *Journal of Otolaryngology-Head and Neck Surgery,* vol. 87, no. 2, pp. 239-245, |1979.

[18] R.|H.|Wilson, R.|McArdle, K.|L.|Watts, and S.|L.|Smith. "The revised speech perception in noise test (R-SPIN) in a multiple signal-to-noise ratio paradigm," *Journal of the American Academy of Audiology,* vol. 23, no. 8, pp. 590-605, 2012.

[19] M.|S.|Sommers, N.|Tye-Murray, and B.|Spehar. "Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults," *Ear and Hearing*, vol. 26, no. 3, pp. 263-275, 2005.

[20] R.|W.|McCreery, M.|Spratford, B.|Kirby, and M.|Brennan. "Individual differences in language and working memory affect children's speech recognition in noise," *International Journal of Audiology,* vol. 56, no. 5, pp. 306-315, 2017.

[21] Apple Final Cut Pro X version 10.4. Cupertino, California: Apple Inc., 2017.

[22] J.|M.|Saragih, S.|Lucey, and J.|E.|Cohn. "Deformable model fitting by regularized landmark mean-shift," *International Journal Computational Vision*, vol. 91, no. 2, pp. 200-215, 2011.

[23] MATLAB version 9.6.0. Natick, Massachusetts: The MathWorks Inc., 2019.

[24] Adobe Audition version 11.1.0. San Jose, California: Adobe Inc., 2018.

[25] MathWorks, *Signal Processing Toolbox (R*elease *2019a*). Natick, Massachusetts: The MathWorks, Inc., 2019.

[26] K.|W.|Grant and P.|F.|Seitz. "Measures of auditory-visual integration in nonsense syllables and sentences," *Journal of the Acoustical Society of America,* vol. 104, no. 4, pp. 2438-2450, 1998.

[27] K.|G.|Munhall, P.|Gribble, L.|Sacco, and M.|Ward. "Temporal constraints on the McGurk effect," *Perception and Psychophysics,* vol. 58, no. 3, pp. 351-362, 1996.

[28] Max 7. Version 7.3.5. San Francisco, California: Cycling '74, 2018.

[29] American Speech-Language-Hearing Association. "Guidelines for determining threshold levels for speech," *American Speech-Language-Hearing Association,* vol. 134, no. 6, pp. 85-89, 1998.

[30] D.|N.|Kalikow, K.|N.|Stevens, and L.|L.|Elliott. "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *Journal of the Acoustical Society of America,* vol. 61, no. 5, pp. 1337–1351, 1977.

[31] H.|Levitt. "Transformed up-down methods in psychoacoustics," *Journal of the Acoustical Society of America,* vol. 49, no. 2B, pp. 467-477, 1971.

[32] D.|Bates, M.|Maechler, B.|Bolker, and S.|Walker. *lme4: Linear mixed effects models using Eigen and S4.* R package version 1.1-8, 2015.

[33] R Core Team, *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/, 2017.

[34] S.|ten Oever, C.|E.|Schroeder, D.|Poeppel, N.|van Atteveldt, and E.|Zion-Golumbic. "Rhythmicity and cross-modal temporal cues facilitate detection," *Neuropsychologia*, vol. 63, pp. 43-50, 2014.

[35] C.|E.|Schroeder and J.|J.|Foxe. "Multisensory contributions to low-level 'unisensory' processing," *Current Opinion in Neurobiology,* vol. 15, no. 4, pp. 454-458, 2005.

[36] C.|E.|Schroeder, P.|Lakatos, Y.|Kajikawa, S.|Partan, and A.|Puce. "Neuronal oscillations and visual amplification of speech," *Trends in Cognitive Science,* vol. 12, no. 3, pp. 106-113, 2008.

[37] E.|Zion-Golumbic, G.|B.|Cogan, C.|E.|Schroeder, and D.|Poeppel. "Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party"," *Journal of Neuroscience,* vol. 33, no. 4, pp. 1417-1426, 2013.

[38] J.|Kim, V.|Aubanel, and C.|Davis. "Effect of auditory and visual signal availability on speech perception," in *Proceedings of the 18th International Conference on Phonetic Science, August 10-14, Glasgow, UK,* 2019.

[39] S.|A.|Wood. "Relative benefits of linear analogue and advanced digital hearing aids," *International Journal of Audiology,* vol. 43, no. 3, pp. 144-155, 2004.

[40] L.|M.|Jenstad, R.|C.|Seewald, L.|E.|Cornelisse, and J.|Shantz. "Comparison of linear gain and wide dynamic range compression hearing aid circuits: aided speech perception measures," *Ear and Hearing,* vol. 20, no. 2., pp. 117-126, 1999.

[41] P.|Souza. "Effects of compression on speech acoustics, intelligibility, and sound quality." *Trends in Amplification,* vol. 6, no. 4, pp. 131-165, 2002.