3-10-2020

Visual analog of the acoustic amplitude envelope benefits speech perception in noise

Yi Yuan
San Jose State University, yi.yuan@sjsu.edu

Ratree Wayland
University of Florida

Yonghee Oh
University of Florida

Recommended Citation

Yi Yuan, Ratree Wayland, and Yonghee Oh. "Visual analog of the acoustic amplitude envelope benefits speech perception in noise" *The Journal of the Acoustical Society of America* (2020): 246-251. https://doi.org/10.1121/10.0000737

The Journal of the Acoustical Society of America

MARCH 10 2020

Visual analog of the acoustic amplitude envelope benefits speech perception in noise

Yi Yuan; Ratree Wayland; Yonghee Oh

J. Acoust. Soc. Am. 147, EL246–EL251 (2020)
https://doi.org/10.1121/10.0000737

ASA
ACOUSTICAL SOCIETY OF AMERICA

JASA EXPRESS LETTERS

https://doi.org/10.1121/10.0000737

Visual analog of the acoustic amplitude envelope benefits speech perception in noise

Yi Yuan,[1,a] Ratree Wayland,[2] and Yonghee Oh[1]

[1]Department of Speech, Language, and Hearing Sciences, University of Florida, Gainesville, Florida 32610, USA

[2]Department of Linguistics, University of Florida, Gainesville, Florida 32611, USA

yiyuan56@ufl.edu, ratree@ufl.edu, yoh@phhp.ufl.edu

Abstract: The nature of the visual input that integrates with the audio signal to yield speech processing advantages remains controversial. This study tests the hypothesis that the information extracted for audiovisual integration includes co-occurring suprasegmental dynamic changes in the acoustic and visual signal. English sentences embedded in multi-talker babble noise were presented to native English listeners in audio-only and audiovisual modalities. A significant intelligibility enhancement with the visual analogs congruent to the acoustic amplitude envelopes was observed. These results suggest that dynamic visual modulation provides speech rhythmic information that can be integrated online with the audio signal to enhance speech intelligibility.

[Editor: Martin Cooke] Pages: EL246–EL251

Received: 14 October 2019 **Accepted:** 29 January 2020 **Published Online:** 10 March 2020

1. Introduction

Speech perception often occurs in the presence of various background noises, reverberation, and competing voices, which masks target utterances. This can lead to difficulty in speech perception for normal listeners, and even more pronounced difficulty for hearing-impaired listeners (Rabbitt, 1968; Helfer and Wilber, 1990; Pichora-Fuller and Singh, 2006). Numerous studies have shown that speech intelligibility can be enhanced under audiovisual (AV) conditions (Bernstein *et al.*, 2003; Grant, 2001; Grant and Seitz, 2000; Sumby and Pollack, 1954). Adding visual information such as lip movements substantially improves speech intelligibility at low speech-to-noise ratios (SNRs; Sumby and Pollack, 1954). The benefits of different types of visual information, including head-motion animated faces (Yehia *et al.*, 2002), point-light displays (Jaekl *et al.*, 2015), and degraded facial videos (Alsius *et al.*, 2016), have been investigated.

However, the nature of the visual and audio input information to be integrated in AV speech perception remains the subject of debate. According to Rosenblum's "amodal" or "modality-neutral" model, the acoustic and optic signals take on a similar form because speech articulatory gestures affect them both in a similar way. The integration of the commonly shared, higher-order information from both signals is "simply a consequence and property of the input information itself" (Rosenblum, 2008, p. 406). According to this model, auditory and visual information are functionally the same because of their intimate link to articulatory gestures (Rosenblum, 2008). Another model, which was proposed by Berthommier (2004), suggested that the speech sounds and related mouth movements are integrated at a pre-phonetic level. AV speech perception is based on the product between temporal correlated audio envelopes and visually-predicted amplitude envelopes of acoustic speech.

The hypothesis of this study is that the information extracted for AV integration includes co-occurring suprasegmental dynamic changes in the acoustic and visual signals. The current set of experiments was designed to investigate whether listeners can benefit from dynamic temporal visual information that is correlated with the auditory speech signal but does not contain specific articulatory information.

As a departure from previous visual speech processing studies, this study examines the effects of an abstract visual analog of the acoustic speech amplitude envelope in AV processing. The amplitude envelope of acoustic speech contains temporal structures that would imply the segmentations in speech and relate to rhythmic structures (Summerfield, 1992; Grant and Seitz, 2000). A visual analog of the acoustic speech amplitude envelope, rather than facial and lip movements, was used as visual cues in two experimental settings proposed in this study: (1) congruent AV temporal coherence, and (2) incongruent AV temporal coherence. In each experiment,

[a]Author to whom correspondence should be addressed.

https://doi.org/10.1121/10.0000737

target sentences were presented in audio-only and AV conditions. The number of words accurately recognized in the two modes of presentation was compared in order to quantify the benefits of AV speech processing benefits. To the extent that listeners showed a benefit for the AV condition, it provides evidence for the hypothesis that listeners can utilize dynamic temporal visual information that is correlated with the auditory speech signal but does not contain any information about specific phonemic articulation.

2. General methods

2.1 Subjects

These experiments were conducted according to the guidelines for the protection of human subjects as set forth by the Institutional Review Boards of the University of Florida. Even though there are no gender differences observed in AV benefit for young adults (Alm and Behne, 2015), gender effect was not included in our experimental design. Eighty-three female adult subjects (average age 21.2 ± 4.6 years old) were recruited. All subjects had normal hearing sensitivity (air conduction thresholds ≤25 dB hearing level), and were native, monolingual English speakers.

2.2 Stimulus materials

The speech material for all experiments consisted of 46 of the Harvard sentences (IEEE, 1969) with 23 sentences spoken by one male native English speaker, and the other 23 sentences spoken by one female native English speaker. Six sentences were chosen for the practice section and 40 sentences were chosen for the measurement section. Across different experiments, the same 20 target sentences were always chosen for the audio-only condition and another same set of 20 sentences for the AV condition. Target sentences were embedded within background multi-talker babble noise, with a random amount of duration of the noise (100–400 ms) added before and after the target sentences. The babble noise was prerecorded in a conference room with eight people reading different sentences from the Harvard sentence collection (non-overlapping with the target sentences).

Behavioral piloting data were used to select appropriate SNRs for testing. Eight separately recruited subjects were tested on 30 sentences at SNRs of −1, −3, and −5 dB from both male and female speakers presented in the auditory-only condition. Participants were asked to type down the target sentences they heard, and the number of words accurately identified was calculated as a percentage of the total number of words in the target sentence. The results indicate that an SNR of −3 dB for both female [mean = 77.03%, standard deviation (SD) = 20.64%] and male (mean = 62.11%, SD = 28.64%) speakers yielded an appropriate level of performance to avoid ceiling and floor effects.

The amplitude envelope for each recorded sentence was extracted using the Hilbert function in MATLAB (R2018b, the MathWorks, Natick, MA). These extracted envelopes were then low-pass filtered with a cutoff frequency of 30 Hz. Figure 1(A) shows the waveform of a sample speech sentence "Rice is often served in round bowls." Figure 1(B) shows the extracted temporal envelope of the same sentence.

The corresponding visual stimulus was a colored sphere that changed volume in synchrony with the acoustic amplitude envelope of each sentence. Two AV temporal coherences were applied: congruent (experiment 1) and incongruent (experiment 2). The script for the creation of the visual stimulus was also created in MATLAB. The videos were rendered into 1280*720- pixel movies at 30 frames/s. All audio stimuli were sampled at 44 100 Hz and root-mean-square (rms) matched through MATLAB at 75 dB sound pressure level for presentation. Figure 1(C) shows an example of the sphere-shaped visual stimuli with the congruent AV temporal coherence at frames 16 ($t = 0.53$ s), 32 ($t = 1.07$ s), 48 ($t = 1.6$ s), and 64 ($t = 2.13$ s). The audio and video files were combined and exported to an AV format using AVS video editor (Online Media Technologies Ltd., London, UK).

2.3 Procedure

Data collection are conducted in a double-walled, sound-attenuated booth. All stimuli were presented through E-prime (version 2.0, Psychology Software Tools). Audio stimuli were presented through HP 380 Pro Sennheiser headphones (Sennheiser), and the visual stimuli were shown on a 17-in.- monitor (1707 FP, DELL). For practice, six AV stimuli were presented with no background babble noise. Participants were asked to focus on the monitor and type the sentence that they heard into an input text window which showed up at the end of the stimulus presentation. Feedback in terms of a written presentation of the correct sentence was included only during this short practice to demonstrate the task. For the testing session, 20 audio-only files and 20 AV files were presented with the recorded background babble noise. Mode of presentation (audio-only versus AV) was randomized from trial to trial. Participants were instructed to keep watching the monitor until the stimulus presentation finished and then asked to type the sentence as they heard it. The percentage of words accurately identified was calculated. The scoring was done separately

JASA EXPRESS LETTERS

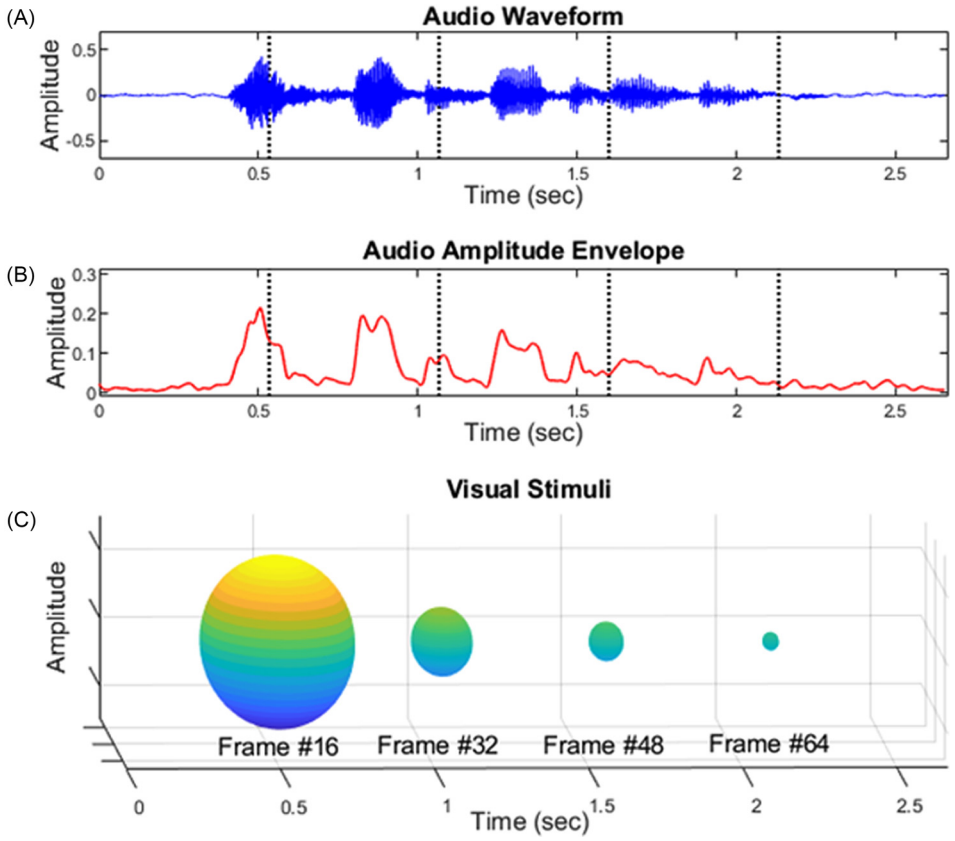

Fig. 1. (Color online) Schematic representation of the stimuli used in this study. (A) The speech waveform for the sentence "Rice is often served in round bowls." (B) Extracted temporal envelope with the 30-Hz cutoff frequency. (C) Sphere-shaped visual stimuli positively synchronized with the acoustic speech amplitude envelope at frame numbers, 16, 32, 48, and 64. The frame rate is 30 frames/s. Vertical dotted lines indicate that sampling points between audio and visual stimuli are in sync.

and cross-checked by two trained research assistants, based on the same scoring rules (for instance, missing articles and wrong spelling were counted as errors). The dependent variable was the percentage of correctly identified words from all target sentences presented. The mean identification accuracy scores in the two conditions (audio-only versus AV) were compared using a linear mixed effects model (LMM) for all experiments.

3. Experiment 1: Congruent AV temporal coherence

According to studies from Grant (2001) and Grant and Seitz (2000), AV speech perception relies on the detection of positive correlations between lip movements and acoustic amplitude. In addition, Summerfield (1992) hypothesized that speech signals and lip movements share the same spatial and temporal cues. Opening of the lip area relates to the opening of the vocal tract. This means that larger amplitude peaks will be observed in acoustic signals. Experiment 1 aims to test our hypothesis that the essence of amplitude envelope information is to convey rhythmic pattern of speech signals rather than gestural information. Congruent AV temporal coherence alone improves speech intelligibility. Since the rhythmic pattern remains no matter if the visual sphere volume and the amplitude modulation is positively or negatively correlated, we hypothesized that AV speech benefits will be observed under both conditions. Two subset tests were included in experiment 1. In the Congruent-Positive (CP) test, the visual stimulus in the AV condition was a colored sphere that increased and decreased in volume synchronized positively with the amplitude envelope of the target sentences [see Fig. 1(C)]. Twenty listeners were presented 40 spoken sentences in background babble noise either in audio-only ($n = 20$) or AV ($n = 20$) condition. In the Congruent-Negative (CN) test, the visual stimulus was synchronized negatively (i.e., higher acoustic speech amplitude = smaller sphere volumes) with the amplitude envelope of the target sentences. Another 20 listeners participated in this test with 20 audio-only stimuli and 20 AV stimuli.

3.1 Results and discussion

The accuracy of the responses was analyzed using a logistic LMM in R (version 3.1.3, R Core Team, 2014) using lmer in the package lme4 (version 1.1-7, Bates *et al.*, 2014). In the CP test,

https://doi.org/10.1121/10.0000737

Table 1. A LMM was applied to analyze the accuracy of the responses. Accuracy was used as the dependent variable, and speech perception condition (audio-only and AV) was added as the fixed effects. Subject and sentence were included as the random effects. ***$P < 0.001$; **$P < 0.01$, *$P < 0.05$.

Experiment	Tests	Effect	Estimate	SE	*t*-value	*P*-value
Exp. 1 (Congruent)	Positive	Intercept (Audio-only)	0.6804	0.0191	35.614	0.000***
		AV	0.0341	0.0185	1.843	0.0657
	Negative	Intercept (Audio-only)	0.61885	0.0178	34.76	0.000***
		AV	0.05325	0.0198	2.690	0.0073**
Exp. 2 (Incongruent)	Mismatched	Intercept (Audio-only)	0.6885	0.02689	25.609	0.000***
		AV	0.0606	0.03802	1.593	0.120
	Randomized	Intercept (Audio-only)	0.59751	0.04566	13.087	0.000***
		AV	0.04872	0.06457	0.755	0.455

condition (audio-only/AV) was included as fixed effects. A maximal random effect structure was used which included by-subject and by-item (target sentences) random intercepts, with the fixed effect and their interactions as by-subject and by-item random slopes. There is a marginal significant effect with AV condition [estimate: 0.0341, standard error (SE): 0.0185, T: 1.843, $p = 0.0657$] (as shown in Table 1): responses for the AV condition were more accurate than for the audio-only condition [mean proportion of accurate responses in the AV condition: 0.71 (SD 0.07); audio-only condition: 0.68 (SD 0.091), Cohen's $d = 0.41$]. In the CN test, the model included condition (audio-only/AV) as fixed effect, by-subject as random intercept. There is a significant effect with AV condition [estimate: 0.05325, SE: 0.0198, T: 2.690, $p = 0.007$]: a significantly higher number of words were recognized in the AV condition (mean: 0.67, SD: 0.07) than in the audio-only condition (mean: 0.62, SD: 0.07).

In both tests, significant improvements in accuracy in the AV condition versus the audio-only condition was obtained even though no visual representation of articulatory information was available. These results suggest that visual representations of the amplitude envelope can be integrated online during speech perception. To be more specific, visual speech perception benefits could result from AV integration of time-varying, rhythmic patterns (for instance, information of onset and offset of the syllables) that co-occur in heard and seen speech. Note that participants were not informed about the relationship of the visual and audio stimuli. They also had no experience with this bimodal correlation as opposed to the extensive experience that most listeners have with face-to-face speech.

4. Experiment 2: Incongruent AV temporal coherence

Results of experiment 1 revealed that the congruent visual analog of the acoustic amplitude envelope improved its intelligibility in noise by facilitating AV integration. Experiment 2 was designed to see how coupled the visual dynamic information has to be with the amplitude envelope in order to obtain the AV benefits for speech perception in noise. In experiment 2, visual analogs of the amplitude envelopes did not correspond to the audio signals. Two subset tests were included: (1) the Incongruent-Mismatched (IM) test, where the target audio signal was mismatched with visual analog, a ball changing volume based on the amplitude envelope from another sentence instead of combining with visual analog of its own amplitude envelope. (2) The Incongruent-Randomized (IR) test, where the visual stimulus was a ball generated by MATLAB based on randomized modulation rate (2–100 Hz) and modulation depth (0%–100%), which is uncorrelated with any of the target sentences' amplitude envelopes. We hypothesized that the incongruence between visual and audio signals would disrupt the AV integration process. It is also possible that participants would entirely ignore this visual cue and only rely on audio signals. In either case, no intelligibility enhancement would be expected.

4.1 Results and discussion

We ran linear mixed model analysis in R (version 3.1.3, R Core Team, 2014) using the lme4 package (version 1.1–7, Bates *et al.*, 2014) to analyze the accuracy of responses. In the mismatched test, we included accuracy as the dependent variable and added fixed effects of speech perception condition (audio-only and AV). We included by-subject and by-sentence as random effects. Significance was calculated using the lmer Test package (Kuznetsova *et al.*, 2017). The model specification was as follows: lmer (Accuracy ~ Condition + (1| Subject) + (1| Sentence). There was no significant effect with AV condition [estimate: 0.06055, SE: 0.03802, T: 1.593, $p = 0.120$] (as shown in Table 1). In the randomized test, the model included condition as fixed

effect and by-sentence as random intercept. No significant effect was observed with the AV condition [estimate: 0.04872, SE: 0.06457, T: 0.755, $p = 0.455$]. Both results indicated that responses between the AV condition and audio-only condition were not different. These findings are consistent with our prediction. In this control condition, incongruent visual stimuli did not provide beneficial cues for AV integration.

5. General discussion and conclusion

Visual information is beneficial only when it is integrated with audio information in speech perception. Numerous studies have demonstrated that visually presenting a talker's face can lower the speech detection threshold and increase perceptual accuracy in noisy environments (Bernstein *et al.*, 2004; Grant and Seitz, 2000; Sumby and Pollack, 1954). These results have been interpreted to support gestural theories of speech perception. According to these theories, the objects of speech perception are intended gestures used in speech production (e.g., Liberman and Mattingly, 1985). In contrast, for general auditory and learning approaches, the objects of speech perception are the auditory qualities of the phonetic segments (Diehl and Kluender, 1989).

The current study examined what aspects of visual input are necessary to yield an AV benefit for speech perception in noise. The primary question addressed is which visual characteristics can facilitate speech perception and whether listeners would benefit from the co-occurring suprasegmental temporal dynamic changes in the acoustic and visual signal. As opposed to many classic investigations of AV integration at the level of the phoneme or syllable (e.g., McGurk and McDonald, 1976), the current study focuses on recognition of sentences and integration at the suprasegmental level (Crosse *et al.*, 2016). Visual stimuli were spherical shapes varying in volume synchronously with the amplitude envelope modulation of the target sentence (congruent condition, both positive and negative correlated), a different sentence (IM condition) and a randomized amplitude modulation (IR). A significant increase in the number of words recognized from the audio-only to the AV modalities was observed in both positively and negatively congruent conditions. These results demonstrated that listeners could benefit in speech perception in noise from the arbitrary dynamic visual source that is correlated with the amplitude envelope. Limitations should be noted in the current study. We did not statistically compare across the experiments and "subtests." We only calculated if accuracy rates were significantly higher in the AV condition compared with the audio-only condition within each subtest. Considering the visual amplitude envelope information available in real AV speech is less precisely coupled to the acoustics, what aspects of this dynamic visual source are most important for the AV benefit would be determined in the future experiments.

To generate visual stimulus, a dynamic cue was extracted independent of any phonetic articulatory cue that listeners can use in this study. Despite this impoverished stimulus, listeners can benefit from the visual cue correlated with the amplitude envelope of acoustic speech. It suggests that some of the AV benefits for speech perception in noise may come from a non-phonetic source. This finding is consistent with the general auditory and learning approaches (Diehl *et al.*, 2004) in speech perception rather than the motor theory (Liberman and Mattingly, 1985). According to the motor theory, speech perception in humans is an articulation gesture perception. Observing lip movements elicits a motor plan in the listeners which could be applied to produce the observed movements (Campbell, 2008; Macdonald and McGurk, 1978). Alternatively, general auditory accounts suggest that the visual information is not about articulations, *per se*, but about correlations between the visual and auditory signal (Diehl *et al.*, 2004). In a study with 7.5-month-old infants (Hollich *et al.*, 2005), with synchronized visual-auditory stimuli (talker's face-target words), infants performed better in a segmentation task, which showed benefit of synchronized auditory-visual signals in infants to segregate target in the speech stream. Since AV speech signals unfold over time, we proposed that lip reading ability is essentially a "speech tracking" ability. This ability relies on temporal information drawn from both audio and visual sensory modalities.

In conclusion, AV speech perception as a multisensory integration process relies on salient temporal cues to enhance both speech detection and tracking ability. Amplitude envelope, as a reliable source of temporal cues, can be delivered through different sensory modalities (e.g., visual) to enhance speech intelligibility when auditory perception ability is compromised. Therefore, the results of this study have implications for potential technological enhancements to speech perception above and beyond hearing devices—in particular, the integration of a non-auditory signal.

Acknowledgment

The authors would like to thank Andrew Lotto for his contribution to the whole project. We also thank Beatrice David, Lauren Husney, Sabrina Lee, Yucheng Liu, and Aaron McEnery for their valuable comments which greatly improved the readability and the quality of the paper.

https://doi.org/10.1121/10.0000737

References and links

Alm, M., and Behne, D. (**2015**). "Do gender differences in audiovisual benefit and visual influence in audio-visual speech perception emerge with age?," Frontiers Psychol. **6**, 1014.

Alsius, A., Wayne, R. V., Paré, M., and Munhall, K. G. (**2016**). "High visual resolution matters in audiovisual speech perception, but only for some," Attn., Percept., Psychophys. **78**(5), 1472–1487.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2014**). "Fitting linear mixed-effects models using lme4," preprint arXiv:1406.5823.

Bernstein, L. E., Auer, E. T., and Takayanagi, S. (**2004**). "Auditory speech detection in noise enhanced by lip reading," Speech Commun. **44**(1-4 SPEC. ISS.), 5–18.

Bernstein, L. E., Takayanagi, S., and Auer, E. T. (**2003**). "Enhanced auditory detection with AV speech: Perceptual evidence for speech and non-speech mechanisms," in *International Conference on Audio-visual Speech Processing*.

Berthommier, F. (**2004**). "A phonetically neutral model of the low-level audio-visual interaction," Speech Commun. **44**(1–4), 31–41.

Campbell, R. (**2008**). "The processing of audio-visual speech: Empirical and neural bases," Philos. Trans. R. Soc. B **363**(1493), 1001–1010.

Crosse, M. J., Di Liberto, G. M., and Lalor, E. C. (**2016**). "Eye can hear clearly now: Inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration," J. Neurosci. **36**(38), 9888–9895.

Diehl, R. L., and Kluender, K. R. (**1989**). "On the objects of speech perception," Ecol. Psychol. **1**(2), 121–144.

Diehl, R. L., Lotto, A. J., and Holt, L. L. (**2004**). "Speech perception," Annu. Rev. Psychol. **55**, 149–179.

Grant, K. W. (**2001**). "The effect of speechreading on masked detection thresholds for filtered speech," J. Acoust. Soc. Am. **109**, 2272–2275.

Grant, K. W., and Seitz, P.-F. (**2000**). "The use of visible speech cues for improving auditory detection of spoken sentences," J. Acoust. Soc. Am. **108**(3), 1197–1208.

Helfer, K. S., and Wilber, L. A. (**1990**). "Hearing loss, aging, and speech perception in reverberation and noise," J. Speech, Lang., Hear. Res. **33**(1), 149–155.

Hollich, G., Newman, R. S., and Jusczyk, P. W. (**2005**). "Infants' use of synchronized visual information to separate streams of speech," Child Develop. **76**(3), 598–613.

IEEE (**1969**). "IEEE recommended practice for speech quality measurements" (Institute of Electrical and Electronic Engineers, New York).

Jaekl, P., Pesquita, A., Alsius, A., Munhall, K., and Soto-Faraco, S. (**2015**). "The contribution of dynamic visual cues to audiovisual speech perception," Neuropsychologia **75**, 402–410.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2017**). "lmerTest package: Tests in linear mixed effects models," J. Stat. Software **82**(13), 1–26.

Liberman, A. M., and Mattingly, I. G. (**1985**). "The motor theory of speech perception revised," Cognition **21**(1), 1–36.

Macdonald, J., and McGurk, H. (**1978**). "Visual influences on speech perception processes," Percept. Psychophys. **24**(3), 253–257.

McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature **264**, 746–748.

Pichora-Fuller, M. K., and Singh, G. (**2006**). "Effects of age on auditory and cognitive processing: Implications for hearing aid fitting and audiologic rehabilitation," Trends Ampl. **10**(1), 29–59.

Rabbitt, P. M. (**1968**). "Channel-capacity, intelligibility and immediate memory," Q. J. Exp. Psychol. **20**(3), 241–248.

R Core Team (**2014**). "R: A language and environment for statistical computing" (R Foundation for Statistical Computing, Vienna, Austria), http://www.R-project.org/.

Rosenblum, L. D. (**2008**). "Speech perception as a multimodal phenomenon," Current Direct. Psychol. Sci. **17**(6), 405–409.

Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**(2), 212–215.

Summerfield, Q. (**1992**). "Audio-visual speech perception, lip-reading and artificial stimulation," Philos. Trans. Biol. Sci. **335**(1273), 71–78.

Yehia, H. C., Kuratate, T., and Vatikiotis-Bateson, E. (**2002**). "Linking facial animation, head motion and speech acoustics," J. Phonetics **30**(3), 555–568.