The 14th International Conference on Auditory-Visual Speech Processing
25-26 August 2017, Stockholm, Sweden

Contribution of visual rhythmic information to speech perception in noise

Vincent Aubanel[1], *Cassandra Masters*[2], *Jeesun Kim,*[2], *Chris Davis*[2]

[1]University of Grenoble Alpes, CNRS, GIPSA-lab, Grenoble, France
[2]MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia

vincent.aubanel@gipsa-lab.fr

Abstract

Visual speech information helps listeners perceive speech in noise. The cues underpinning this visual advantage appear to be global and distributed, and previous research hasn't succeeded in pinning down simple dimensions to explain the effect. In this study we focus on the temporal aspects of visual speech cues. In comparison to a baseline of auditory only sentences mixed with noise, we tested the effect of making available a visual speech signal that carries the rhythm of the spoken sentence, through a temporal visual mask function linked to the times of the auditory p-centers, as quantified by stressed syllable onsets. We systematically varied the relative alignment of the peaks of the maximum exposure of visual speech cues with the presumed anchors of sentence rhythm and contrasted these speech cues against an abstract visual condition, whereby the visual signal consisted of a stylised moving curve with its dynamics determined by the mask function. We found that both visual signal types provided a significant benefit to speech recognition in noise, with the speech cues providing the largest benefit. The benefit was largely independent of the amount of delay in relation to the auditory p-centers. Taken together, the results call for further inquiry into temporal dynamics of visual and auditory speech.

Index Terms: Auditory-visual speech, Speech rhythm, Speech in noise

1. Introduction

In noisy environments, seeing the lips of the interlocutor moving in time helps disambiguate distorted and masked speech sounds reaching the listener's ears. But to what extent is the benefit attributable to the shape of the articulators, and to the timing of the gesture, and how much detail is necessary for visual speech cues to produce a benefit? This study aims to broadly characterise the temporal function of the visual advantage in auditory-visual speech perception in noise.

The visual advantage is long attested [1] and has been quantified as leading to an equivalent intensity increase of 11 dB [2]. However, the search to pin down a set of visual cues that could explain this effect has been met with mixed results [3]; with the general conclusion being that the advantage is a global one and the cues are interchangeable and distributed [4]. Moreover, the notion of visemes [5], the equivalent of phonemes for visual speech perception has not been not as productive as expected in explaining visual speech perception processes, [6].

One idea that long been an object of study, is that the relative timing of visual and auditory modalities in speech perception is an important factor in the visual speech benefit. Early gating studies (e.g. [7, 8]) showed that visual speech information, such as anticipatory lip rounding, can be perceived before auditory speech information, leading to the idea that visual speech helps by *predicting* upcoming speech information. This hypothesis has received considerable attention and experimental support [9, 10, 11], and recent neuroimaging studies have found evidence that visual information speeds up the processing of auditory information ([12, 13, 14, 15]).

In line with this research, [16] conducted a cross-language corpus analysis to quantify the visual-auditory alignment in production, concluding in an approximatively 150 ms visual-lead constant. The generality of these results were however questioned by [17] on the basis of methodological issues in measuring visual-auditory events and that the CVC material (relatively infrequent in common auditory-visual speech situations) overrepresented preparatory gestures. Instead, they showed that the timing of auditory-visual events was symmetrical (from 20 ms auditory-lead to 70 ms visual-lead).

Another method for examining the effect of the relative temporal alignment of events in auditory-visual speech streams has been to experimentally manipulate the synchrony between modalities. This research has shown an asymmetric window from 50 ms audio-lead to 200 ms video-lead within which perceptual simultaneity is experienced, with a maximum reached for visual-lead asynchrony values [18]. A parsing mechanism for visual information corresponding to this temporal window has been proposed as a 'visual syllable' [12, 13, 14], not ruling out the fact that asynchronies at a finer time scale could also be detected [19]. Although the results of such manipulation studies are useful in gauging the sensitivity of the perceptual system to the relative timing of visual and auditory speech events, it should be pointed out that this work may be picking up on the ability of the system to adapt to unusual stimulus conditions rather than on what happens with unmanipulated material.

A recent study [20] succeeded in characterising the temporal relationship between auditory and visual modalities without having to desynchronise the auditory and visual signals by using a *classification image* technique [21, 22]. In a McGurk effect scenario [23], the authors identified the information-bearing visual regions that were the main determinant of listeners' responses, (i.e., the time instant of the maximum velocity of the lip aperture movement in producing /aCa/ sentences was the main determinant of the auditory-visual fusion percept), which occurred *before* the auditory consonant. Further, this temporal visual anchoring was independent of introduced asynchronies between auditory and visual modalities. In addition to replicating the visual-lead influence of the visual modality on auditory-visual speech perception, the results showed that *what* is seen is more important that *when* it is seen.

The study by [20] introduced several important issues in the study of the relationship between auditory and visual speech streams: 1) That it is important to consider which instants in the auditory and visual streams 'anchor' the relationship. 2) That it is important to vary the timing of this information. 3) That it

10.21437/AVSP.2017-18

is important to manipulate the type of information that is presented. Here, we address the first of these issues by focussing on the temporal structure of spoken sentences by considering which aspect of the auditory signal are likely to be important in establishing temporal rhythms. In our view, the important event in the perception of rhythmicity in speech are perceptual centers, or p-centers [24, 25], defined as the time instants at which listeners would place a beat to a spoken sentence. In the context of recent models looking at cortical tracking of speech (e.g. [26, 27]), we recently proposed that p-centers, rather than amplitude envelope peaks, could be the support for cortical tracking in noise, showing that isochronous sentences anchored to p-centers were more intelligible than matched anisochronous ones [28]. It has recently been proposed that visual speech could also play a role in guiding cortical entrainment, modifying the phase of cortical oscillations to reach maximum excitability at expected important auditory events [29, 10]. In this context, then, we aimed to test whether there are key times in the unfolding of speech in noise for which seeing the status of visible articulators will be more effective in aiding speech perception.

We also explored to what extent auditory-visual p-center alignment influenced performance by varying the timing of when this information was presented. It is important to note that we did not manipulate the asynchrony between auditory and visual modalities, but instead modified the time instant at which the visual signal is made available. Finally, focussing on the temporal structure also raises the question of whether the temporal information alone could be beneficial. To test this, we contrasted a natural moving face with an abstract visual signal, composed of a stylised moving shape following the same temporal structure. Here, a previous study [30] found that presenting an abstract signal composed of the equivalent of the lip area was not sufficient to provide a speech processing benefit in noise.

The remainder of the paper is organised as follows: in Section 2 we describe how the stimuli were constructed, the experimental design and the data collection. We present an analysis of listeners' responses in Section 3 and discuss the results in Section 4 before concluding in Section 5.

2. Methods

2.1. Stimuli

Auditory-visual recordings of IEEE sentences [31] were taken from the MAVA corpus [32]. A typical sentence of that material consists of five monosyllabic keywords, with its ending mildly predictable from its beginning, as in '*The **latch** on the **back gate needs a nail***' (keywords in bold). The first 165 sentences of the corpus with a minimum of 5 keywords was selected, and an additional 14 sentences were used for practice and catch trials (see section 2.3). Annotation of stressed vs. unstressed syllables were manually checked and auditory p-centers were taken as the onsets of stressed syllables. Table 1 summarises some key features of the test sentences.

Table 1: *Mean (SD) descriptive parameters per sentence (N=165). ISI is the mean inter-stressed syllables interval.*

N words	**N keywords**	**duration (s)**
8.0 (1.2)	5.0 (0.0)	2.2 (0.2)
N syllables	**N stressed syll.**	**ISI (ms)**
9.0 (1.2)	5.1 (0.3)	402 (54)

The visual mask function was defined as 200 ms hanning windows centered around anchor points with a maximum value of 1 at these time points, and 0 elsewhere else. Figure 1 shows a visual mask function with two different alignments for an example sentence.

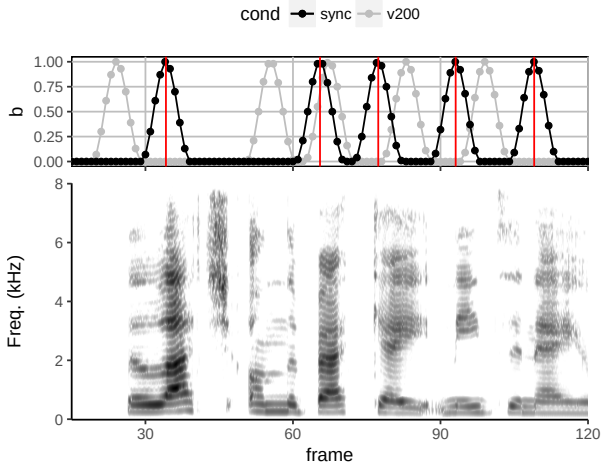

Figure 1: *Visual mask function for the sentence* The latch on the back gate needs a nail. *Top panel: Black line : visual mask function aligned to the auditory p-centers (vertical red lines). Gray line: visual mask function aligned to* -200 *ms prior to auditory p-centers. (Remaining three alignment conditions not shown). Bottom panel: spectrogram.*

In addition to an auditory-only control condition (baseline) that consisted of a static video frame, the **visual modality** consisted of two types of presentation (*natural* and *abstract*).

Video sequences for the *natural* condition were obtained as follows. First, videos were converted to grayscale in order to minimise the general visual impact of varying the transparency. Then, a full opacity baseline value was set on a sentence basis as the mean grayscale value of the video file. Final stimuli were obtained by applying the visual mask frame-by-frame, resulting in full transparency at auditory p-centers and full opacity elsewhere.

Video sequences for the *abstract* condition consisted of a time-varying visual shape representing a stylised impulse, whose height was determined by the visual mask function.

For both the natural and abstract conditions, the **alignment** of the visual mask was explored by varying the location of the visual mask anchor points in relation to the auditory p-centers. Alignment values ranged from -200 to 200 ms in 100 ms steps. Taken altogether, they were 11 conditions: 1 auditory-only control condition (Ctrl.), 5 natural conditions (N_{-200}, N_{-100}, N_0, N_{100}, N_{200}) and 5 abstract conditions (A_{-200}, A_{-100}, A_0, A_{100}, A_{200}). Figure 2 presents a frame-by-frame comparison between natural and abstract conditions for a 200 ms extract of the example sentence.

The auditory signal, common for all video sequences, was obtained by mixing the recording of the spoken sentence with speech shape noise (SSN) of the talker at a signal-to-noise ratio of -3 dB, a value chosen to target around 50% correct responses. SSN was constructed by filtering white noise with 200 LPC coefficients taken from the long-term average speech spectrum computed on a concatenation of all sentence recordings of the talker. RMS energy of sentence-plus-noise mixtures

were individually adjusted to a fixed value of 0.04. Catch trial sequences for non-auditory-only stimuli were constructed by overlaying a red cross on the top right corner of the video frame for the second half of the duration of the sequence. All stimuli were generated offline and played back in a predetermined order (see section 2.3).

2.2. Participants

Forty-two participants were recruited through the Western Sydney University participant pool system and through personal acquaintances. University students received course credit for participation, distant acquaintances received 15 AUD for their participation and close acquaintances received no remuneration. All participants provided informed consent, and all research procedures were approved by the Human Research Ethics Committee of Western Sydney University under the reference H9495. Data from 11 participants were discarded following screening for non-native language (2 participants), hearing screening (2 participants) and performance-based exclusion criteria (7 participants, see section 2.3), leaving 31 participants (7 males) with mean age of 29.2 (SD = 12.5) for further analysis.

2.3. Procedure

Participants were tested individually and were seated in a sound attenuated booth in front of a computer screen, where they were presented with online instructions. The experiment was run on MacBook Pros running Psychtoolbox [33]. Auditory-visual speech-plus-noise mixtures were presented in blocks and participants had to type what they heard. The experiments were self-paced, and participants could take a break after each block. Stimuli were presented over BeyerDynamic DT 770 Pro 80 Ohm closed headphones at a fixed comfortable level.

Sentences were blocked in 11 sets of 15 sentences. The first block was always the auditory-only condition, followed by 5 blocks for one type of the visual modality (natural or abstract) and 5 blocks for the other visual modality conditions. The order of visual modality conditions was counterbalanced between participants, and the order of blocks for different alignment conditions was randomised. Sentences were randomly distributed across the 11 conditions for each participant so that each participant heard each sentence only once and each sentence could be heard in different conditions across participants. Nine practice sentences were presented in groups of three before blocks 1, 2 and 7 as practice, one of which was always a catch trial, and an additional 5 catch trials were distributed randomly in the test blocs. In sum, participants heard 179 sentences, 165 of which were used for scoring. Participants who responded to 40% or more of the catch trials over the course of the experiment were discarded from further analysis. The total duration of the experiment was about 50 min.

3. Results

Participants' responses were automatically scored, discarding non keywords and accounting for common spelling mistakes. Figure 3 shows the mean proportion of correct keyword per condition.

We fitted a generalised linear mixed model to examine the effect of the visual modality and alignment on recognition scores (R package *lme4*, [34]). All conditions were flattened as a single fixed effect factor with 11 levels and intercept for participant and sentence were taken as random effects. Random effects standard deviation for participant and sentence were 0.42 and 1.18 respectively. We then ran simultaneous tests for general linear hypotheses (function `glht()`), specifying a contrast matrix for comparing different conditions and condition groups. Results of the simultaneous multiple comparisons are shown in Table 2.

Table 2: *Output of the simultaneous multiple comparisons on the generalised linear mixed model.*

Comparison	Estim.	Std. Err.	z	p	sig.
Abst., Ctrl.	0.34	0.050	6.9	< 0.001	***
Nat., Ctrl.	0.73	0.050	14.5	< 0.001	***
Nat., Abst.	0.39	0.030	13.0	< 0.001	***
A_{-200}, Abst.	0.03	0.051	0.5	1.000	
A_{-100}, Abst.	0.08	0.051	1.5	0.782	
A_0, Abst.	−0.03	0.052	−0.5	1.000	
A_{100}, Abst.	0.01	0.051	0.1	1.000	
A_{200}, Abst.	−0.08	0.051	−1.6	0.713	
N_{-200}, Nat.	−0.08	0.053	−1.5	0.779	
N_{-100}, Nat.	0.10	0.054	1.8	0.544	
N_0, Nat.	−0.01	0.053	−0.2	1.000	
N_{100}, Nat.	−0.01	0.053	−0.3	1.000	
N_{200}, Nat.	0.01	0.053	0.1	1.000	

Table 2 confirms what is visually apparent from Figure 3. Keywords in all abstract conditions were significantly better identified than keywords in the auditory-only control condition, keywords in the natural conditions were significantly better identified than keywords in the control condition, and keywords in the natural conditions were significantly better identified than keywords in the abstract condition (Table 2, row 1, 2 and 3 respectively). Further comparisons between performace in individual natural [resp. abstract] conditions and all other natural [resp. abstract] conditions did not result in significant differences (Table 2, remaining rows).

In summary, while the type of the visual modality had a clear influence on the recognition scores, the temporal alignment of the visual cue relative to the auditory p-center did not.

4. Discussion

The results showed that the visual benefit was robust across each visual presentation condition and across presentation time. A common result in studies looking at the relative timing between the auditory and visual modalities is the relative tolerance of auditory-visual perception to variations of timing (within a certain range of values, the so called window of integration). Within the window of integration, typically 50 ms audio-lead to 200 ms video-lead, discrimination of stimulus order is difficult, and a visual benefit is found. As proposed by [20], relative timing may not be so important to any benefit, but rather benefit is more to do with the visual informational content of the speech. This proposal is consistent with what was observed in the current study, i.e., there was a visual benefit regardless of when the visual information was presented. The current results are also consistent with a classic study [35] that found a benefit of a visual cue for CVC recognition despite a 1.6 s delay. In fact, the perceptual system seems to be flexible enough to recruit distant information, provided the coherence between modalities has been established. Given the block design of the current study, with a fixed alignment value during each block, it could be the case that global timing recalibration processes are

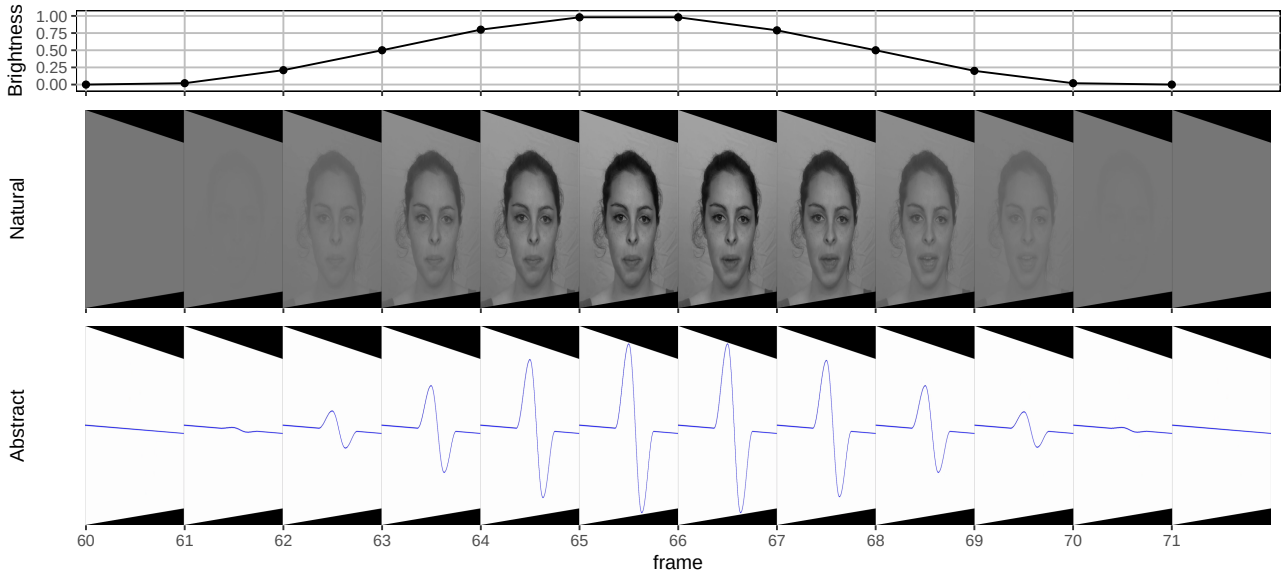

Figure 2: *Frames* $60-71$ *of the example sentence shown in Figure 1, covering approximately the syllable /ba/. Top panel: visual mask function, aligned to the auditory p-center. Middle panel: video frames for the natural condition* N_0*. Note the lip movement associated with the articulation of the /b/. Bottom panel: video frames for the abstract condition* A_0*.*

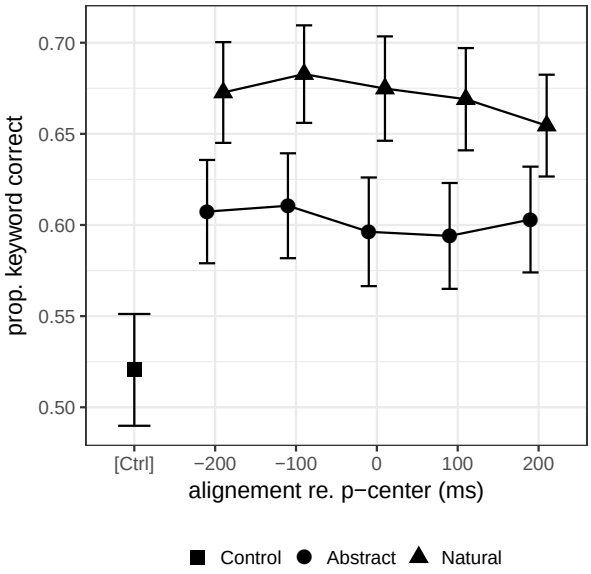

Figure 3: Mean proportion of keyword correct across participants. Errorbars represent 95% confidence intervals ($N = 31$).

at play, whereby the association with the modalities can extend beyond the integration window. It therefore remains an open question as to what the limits of this distant integration may be.

Another possible reason for the relative stability of recognition across different alignment values concerns the type of material used here. Although keywords were only mildly predictable from each other, relevant visual and auditory cues in one location may have favoured recognition in other locations.

A surprising result of the current study was the robust benefit associated with the abstract visual modality, across the different alignment values. Indeed, a previous study did not find a benefit of an abstract representation of the temporal structure to speech perception in noise [30]. Since the only information contained in this visual signal was speech rhythm, the facilitation points to the important role of p-centers in auditory-visual perception. The facilitation also suggests that in the natural condition, there is an additional benefit for speech perception in noise beyond merely seeing the state of the talker's articulators.

5. Conclusions

We showed that making available the timing structure of a spoken utterance is beneficial. The benefit is largely independent of the precise auditory-visual timing alignment. The benefit is even present (although to a lesser degree) when the temporal information is abstract.

6. Acknowledgements

The research leading to these results was partly funded by the Autralian Research Council under grant agreement DP130104447. Author VA also acknowledges support from the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013 Grant Agreement no. 339152, "Speech Unit(e)s", J.-L. Schwartz PI).

7. References

[1] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Am.*, vol. 26, no. 2, pp. 212–215, 1954.

[2] A. MacLeod and Q. Summerfield, "Quantifying the contribution of vision to speech perception in noise," *Brit. J. Audiol.*, vol. 21, no. 2, pp. 131–141, May 1987.

[3] V. Aubanel, C. Davis, and J. Kim, "Explaining the visual and masked-visual advantage in speech perception in noise: The role of visual phonetic cues," in *Proc. of FAAVSP*, Vienna, Austria, 2015.

[4] E. Vatikiotis-Bateson and K. G. Munhall, "Auditory-visual speech processing: Something doesn't add up," in *The Handbook of Speech Production*, M. A. Redford, Ed. John Wiley & Sons, 2015, pp. 178–199.

[5] C. G. Fisher, "Confusions among visually perceived consonants," *J. Speech Lang. Hear. R.*, vol. 11, no. 4, pp. 796–804, 1968.

[6] M. D. Ramage, "Disproving Visemes As The Basic Visual Unit Of Speech," Ph.D. dissertation, Curtin University, Dec. 2013.

[7] M. A. Cathiard, M.-T. Lallouache, S. H. Mohammadi, and C. Abry, "Configurational vs. temporal coherence in audio-visual speech perception," in *Proc. of ICPhS*, Stockholm, Sweden, 1995.

[8] M. A. Cathiard, G. Tiberghien, A. Tseva, M.-T. Lallouache, and P. Escudier, "Visual perception of anticipatory rounding during acoustic pauses: A cross-language study," in *Proc. of ICPhS*, 1991.

[9] K. W. Grant and P.-F. Seitz, "The use of visible speech cues for improving auditory detection of spoken sentences," *J. Acoust. Soc. Am.*, vol. 108, no. 3, p. 1197, 2000.

[10] C. E. Schroeder, P. Lakatos, Y. Kajikawa, S. Partan, and A. Puce, "Neuronal oscillations and visual amplification of speech," *Trends Cogn. Sci.*, vol. 12, no. 3, pp. 106–113, Mar. 2008.

[11] E. M. Z. Golumbic, D. Poeppel, and C. E. Schroeder, "Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective," *Brain Lang.*, vol. 122, no. 3, pp. 151–161, 2012.

[12] L. H. Arnal, B. Morillon, C. A. Kell, and A.-L. Giraud, "Dual neural routing of visual facilitation in speech processing," *The Journal of Neuroscience*, vol. 29, no. 43, pp. 13 445–13 453, 2009.

[13] J. J. Stekelenburg and J. Vroomen, "Neural correlates of multisensory integration of ecologically valid audiovisual events," *Journal of Cognitive Neuroscience*, vol. 19, no. 12, pp. 1964–1973, 2007.

[14] V. van Wassenhove, K. W. Grant, and D. Poeppel, "Visual speech speeds up the neural processing of auditory speech," *P. Natl. Acad. Sci. USA*, vol. 102, no. 4, pp. 1181–1186, Jan. 2005.

[15] T. Paris, J. Kim, and C. Davis, "Visual form predictions facilitate auditory processing at the N1," *Neuroscience*, vol. 343, pp. 157–164, 2017.

[16] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar, "The natural statistics of audiovisual speech," *PLoS Comp. Biol.*, vol. 5, no. 7, p. e1000436, 2009.

[17] J.-L. Schwartz and C. Savariaux, "No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag," *PLoS Comp. Biol.*, vol. 10, no. 7, 2014.

[18] J. Vroomen and M. Keetels, "Perception of intersensory synchrony: a tutorial review," *Attention Perception & Psychophysics*, vol. 72, no. 4, pp. 871–884, 2010.

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

[20] J. H. Venezia, S. M. Thurman, W. Matchin, S. E. George, and G. Hickok, "Timing in audiovisual speech perception: A mini review and new psychophysical data," *Attention Perception & Psychophysics*, vol. 78, no. 2, pp. 583–601, 2016.

[21] A. Ahumada, Jr and J. Lovell, "Stimulus features in signal detection," *J. Acoust. Soc. Am.*, vol. 49, no. 6B, pp. 1751–1756, 1971.

[22] F. Gosselin and P. G. Schyns, "Bubbles: a technique to reveal the use of information in recognition tasks," *Vision Research*, vol. 41, no. 17, pp. 2261–2271, 2001.

[23] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[24] J. Morton, S. Marcus, and C. Frankish, "Perceptual centers (P-centers)," *Psychol. Rev.*, vol. 83, no. 5, p. 405, 1976.

[25] S. K. Scott, "P-centers in speech: An acoustic analysis," Ph.D. dissertation, UCL, London, UK, 1993.

[26] C. E. Schroeder and P. Lakatos, "Low-frequency neuronal oscillations as instruments of sensory selection," *Trends Neurosci.*, vol. 32, no. 1, pp. 9–18, 2009.

[27] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations," *Nat. Neurosci.*, vol. 15, no. 4, pp. 511–517, Apr. 2012.

[28] V. Aubanel, C. Davis, and J. Kim, "Exploring the role of brain oscillations in speech perception in noise: intelligibility of isochronously retimed speech," *Front. Hum. Neurosci.*, vol. 10, no. 430, Aug. 2016.

[29] C. Kayser, C. I. Petkov, and N. K. Logothetis, "Visual modulation of neurons in auditory cortex," *Cereb. Cortex*, vol. 18, no. 7, pp. 1560–1574, 2008.

[30] J.-L. Schwartz, F. Berthommier, and C. Savariaux, "Seeing to hear better: evidence for early audio-visual interactions in speech identification," *Cognition*, vol. 93, no. 2, pp. B69–B78, Sep. 2004.

[31] E. H. Rothauser, W. D. Chapman, N. Guttman, M. H. L. Hecker, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, M. Weistock, V. E. McGee, U. P. Pachl, and W. D. Voiers, "IEEE Recommended practice for speech quality measurements," *IEEE Trans. Audio Acoust.*, pp. 225–246, 1969.

[32] V. Aubanel, C. Davis, and J. Kim, "MAVA (MARCS Auditory-Visual Australian recordings of IEEE sentences)," http://alveo.edu.au/collections/mava/, 2016.

[33] M. Kleiner, D. Brainard, D. Pelli, A. Ingling, R. Murray, and C. Broussard, "What's new in Psychtoolbox-3," *Perception*, vol. 36, no. 14, p. 1, 2007.

[34] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting Linear Mixed-Effects Models Using lme4 ," *Journal of Statistical Software*, pp. 1–48, Oct. 2015.

[35] R. Campbell and B. Dodd, "Hearing by eye," *Quarterly Journal of Experimental Psychology*, vol. 32, no. 1, pp. 85–99, 1980.