Visual-tactile integration in speech perception: Evidence for modality neutral speech primitives

Katie Bicevskis
Department of Linguistics, University of British Columbia, Totem Field Studios (Main Department), 2613 West Mall, Vancouver, British Columbia V6T 1Z4, Canada

Donald Derrick[a)]
University of Canterbury, NZILBB, Private Bag 4800, Christchurch 8140, New Zealand

Bryan Gick[b)]
Department of Linguistics, University of British Columbia, Totem Field Studios (Main Department), 2613 West Mall, Vancouver, British Columbia V6T 1Z4, Canada

(Received 12 February 2016; revised 26 September 2016; accepted 10 October 2016; published online 9 November 2016)

Audio-visual [McGurk and MacDonald (1976). Nature **264**, 746–748] and audio-tactile [Gick and Derrick (2009). Nature **462**(7272), 502–504] speech stimuli enhance speech perception over audio stimuli alone. In addition, multimodal speech stimuli form an asymmetric window of integration that is consistent with the relative speeds of the various signals [Munhall, Gribble, Sacco, and Ward (1996). Percept. Psychophys. **58**(3), 351–362; Gick, Ikegami, and Derrick (2010). J. Acoust. Soc. Am. **128**(5), EL342–EL346]. In this experiment, participants were presented video of faces producing /pa/ and /ba/ syllables, both alone and with air puffs occurring synchronously and at different timings up to 300 ms before and after the stop release. Perceivers were asked to identify the syllable they perceived, and were more likely to respond that they perceived /pa/ when air puffs were present, with asymmetrical preference for puffs following the video signal—consistent with the relative speeds of visual and air puff signals. The results demonstrate that visual-tactile integration of speech perception occurs much as it does with audio-visual and audio-tactile stimuli. This finding contributes to the understanding of multimodal speech perception, lending support to the idea that speech is not perceived as an audio signal that is supplemented by information from other modes, but rather that primitives of speech perception are, in principle, modality neutral.
 [http://dx.doi.org/10.1121/1.4965968]

[MSS] Pages: 3531–3539

I. INTRODUCTION

Research over the last half century has demonstrated the multimodal nature of speech production, primarily focusing on the auditory and visual modalities. Evidence for the importance of the visual mode is found in visual enhancement of speech perception (Sumby and Pollack, 1954), as well as the well-known perceptual speech illusion, the McGurk effect (McGurk and MacDonald, 1976), where the presentation of incongruent auditory and visual speech stimuli results in an integrated percept. The effect of multimodal integration is robust; unlike many other illusions it is still maintained even when the perceiver is aware of what is happening. Further, it does not require synchrony of stimuli to be effective (Munhall *et al.*, 1996; van Wassenhove *et al.*, 2007). Munhall *et al.* (1996) tested asynchronous stimuli ranging from −360 ms (where the audio stimulus precedes the visual stimulus) to 360 ms (where the visual precedes the audio) in increments of 60 ms and found that the effect is maintained across a range of temporal asynchronies ranging from synchronous to 180 ms. This range of asynchronies may be thought of as a *window of multimodal integration*. When the visual stimulus leads the audio by more than 180 ms, the rate of integration significantly declines. However, when the audio stimulus precedes the visual by just 60 ms, the rate of integration also significantly declines, demonstrating that the asynchronous durations over which the effect is maintained are asymmetric. This may be explained by appealing to properties of the natural world, i.e., as light travels faster than sound, people have experience perceiving events where the visual information is received before the audio. An extreme example of this is the perception of thunder and lightning. Although these are simultaneous events, the lightning is visible moments before the thunder is audible. Munhall *et al.* (1996) argue that people experience more subtle asynchronous events in daily life and through experience come to perceive the slightly asynchronous visual and auditory components of an event as being simultaneous.

A growing body of research (Alcorn, 1932; Fowler and Dekle, 1991), has shown that the tactile mode also contributes information to the speech stream. Gick and Derrick (2009) showed that when participants were presented with audio stimuli ambiguous between /pa/ and /ba/, they reported hearing more /pa/ stimuli when the audio stimulus was presented simultaneously with a puff of air to the neck or hand; the same effect has been replicated using puffs at the ankle (Derrick and Gick, 2013). When light taps, generated from a solenoid tapping system, were used instead of puffs of air,

[a)]Electronic mail: donald.derrick@gmail.com
[b)]Also at: Haskins Laboratories, New Haven, CT 06511, USA.

the taps had no significant effect on speech perception (see supplementary materials in Gick and Derrick, 2009). This difference in results demonstrates that the participants were not simply responding to general tactile cues, nor was the result generated from increased attention. Instead, these results suggest that speakers integrate the aero-tactile information normally produced during speech.

Following on the question of temporal asynchrony, Gick *et al.* (2010) found that, similar to audio-visual speech events, participants integrate audio-tactile speech over a range of asynchronies. When presented with stimulus asynchronies ranging from −300 ms (tactile stimulus preceding audio stimulus) to 300 ms (audio stimulus preceding tactile stimulus) participants exhibited a window of multimodal integration between −50 and 200 ms. Again the effect is asymmetric in a direction that suggests adherence to the relative speeds of physical signals; as sound travels faster than the pressure front of airflow in speech, integration is more likely to occur when the audio precedes the tactile stimulus.

Despite the substantial literature on multimodal integration in speech, there remains a lack of research into visual-tactile integration in speech perception, which is not surprising considering that auditory information is usually treated as primary in speech (see Diehl *et al.*, 2004). Gick *et al.* (2008) examined the influence of tactile information on visual speech perception using the Tadoma method (Alcorn, 1932), a speech comprehension technique whereby perceivers place a hand in a specified position over the mouth and jaw of a speaker in order to perceive tactile speech information. They found that some untrained perceivers' perception of VCV syllables improved by around ten per cent when they felt the speaker's face whilst watching them silently speak, as opposed to when they had access to only the visual speech information. Research into multimodal speech perception thus shows that perceptual integration can occur with audio-visual, audio-(aero)tactile, and visual-tactile modality combinations. For audio-visual and audio-(aero)tactile speech integration, it also reveals that there is a general temporal window over which this integration is likely to occur and that this window of multimodal integration is asymmetric in a direction which reflects the relative speeds of certain physical properties. Currently unexplored is the potential for individuals to integrate speech information from the visual and aero-tactile modes in the absence of the original auditory speech signal. If integration involving this modality combination does occur, it is as yet unclear whether visual-(aero)tactile speech perception obeys the same principles in terms of temporal window properties as observed for audio-visual and audio-(aero)tactile combinations of speech information. We would expect a similar asymmetric window of integration, but with longer tails as the visual cues for stop release (cheeks puffing, lips opening) are less abrupt than the auditory cues. Since both the airflow at any distance and the visual cues are less abrupt, we expect the window of integration to be similarly less abrupt. If visual-(aero)tactile speech integration were to take place, this finding would contribute new knowledge to our understanding of multimodal speech perception. It would lend support to the idea that speech is not perceived as an audio signal, which is supplemented by information from other modes, but rather that primitives of speech perception are, in principle, modality neutral (Fowler, 2004; Rosenblum *et al.*, 2005; Rosenblum, 2008). Such results would provide behavioral corroboration for neuroimaging results that have long supported a more deeply multimodal view of speech perception (e.g., Calvert *et al.*, 1997). In contrast, Bernstein and Liebenthal (2014) argue that there are visual system specific representations of speech. They argue that these are not just representations of facial motions, but representations of such motion acting in the capacity of speech, and that these can be found within higher-level vision brain areas.

Note that when we here speak of speech as modality neutral, we are not denying the existence of such brain structures. Behavioural research alone cannot tell us when or how multimodal integration operates. Instead, research like that of Sumby and Pollack (1954) shows that in audiovisual speech perception, the two signals provide information, and that if the audio signal-to-noise ratio (SNR) is below about −12 db, the visual signal will provide more information about speech than the audio signal will—and so it massively improves speech perception. By this understanding, no mode is necessarily primary, but instead context and signal strength always apply. Similarly, Auer and Bernstein (1997) and other speech-reading studies show that while lip/face reading is difficult and prone to allowing more signal ambiguity, it alone can be used for highly accurate speech perception. We seek to strengthen this argument from behavioral evidence by showing that cross-modal speech perception integration can occur without an audio signal at all.

The present study investigates the effects of aero-tactile information on visual speech perception of syllables with labial onsets, in the absence of audible speech information. Previous research has shown that English bilabial stops can be considered instances of a single viseme (Fisher, 1968), therefore without the addition of any other speech information, they should not be distinguishable solely through the visual modality (though see an alternative analysis presented in Abel *et al.*, 2011). Nevertheless, there are possible reasons why this might not be the case. In all of the previous audio-aero-tactile research, there was a bias toward /ba/ perception in the audio-only condition (Gick and Derrick, 2009; Gick *et al.*, 2010; Derrick and Gick, 2013). On the other hand, /p/ (2.25% of phones) is more common than /b/ (1.65%) in American English speech (Hayden, 1950).

Based on prior research into multimodal integration, and extending these findings to visual-tactile modality combinations, we make the following hypotheses and predictions:

We hypothesize that tactile speech information contributes to the speech signal in a comparable manner to that of auditory and visual speech information, such that the available information from any modality combination is integrated as part of an overall speech percept. The temporal relationship between tactile stimuli and stimuli from other modalities need not be completely synchronous for this integration to occur, with integration commonly occurring over a temporal window of asynchronies. We expect this window of multimodal integration to be asymmetric because of the differing speeds of ambient physical signals in the world. Considering visual and tactile stimuli, the speed of light is

faster than the speed of airflow, therefore visual-tactile speech stimuli in which the visual stimulus precedes the tactile stimulus is more likely to be integrated than when the tactile stimulus precedes the visual. Light travels faster than sound, and airflow slower, so that we should expect an effect between light and airflow to be substantial. Regarding the current experimental study, the following predictions are made based on these hypotheses:

Following the notion that /b/ and /p/ are considered to be instances of a single viseme (Fisher, 1968), it is first predicted that:

Prediction 1: Participants will perform at chance levels when presented with visual-only /pa/ and /ba/ stimuli.

Previous research into audio-visual (McGurk and MacDonald, 1976) and audio-tactile (Gick and Derrick, 2009) speech perception has demonstrated perceptual integration of speech information from each modality. Assuming that this is evidence for multimodal speech rather than bimodal speech specific to audio-visual and audio-tactile combinations, a second and primary prediction is the following:

Prediction 2: Participants will give more /pa/ responses when presented with synchronous visual and tactile (in the form of air puffs on the skin) stimuli than when visual-only stimuli are presented. Increased /pa/ responses would indicate that participants are integrating the tactile stimuli as perceived aspiration.

Further, previous research into both audio-visual (Munhall *et al.*, 1996; van Wassenhove *et al.*, 2007) and audio-tactile (Gick *et al.*, 2010) speech perception has found that there is a window of asynchronies over which integration is maintained, with stimuli more likely to be integrated when the asynchrony is closer to 0 ms (synchronous). A third prediction is therefore:

Prediction 3 a: Participants will give more /pa/ responses when the stimulus onset asynchrony (SOA) is smaller (closer to synchronous in either direction). Increased /pa/ responses would again indicate that participants are integrating the tactile stimulus as perceived aspiration.

Based on findings related to the physical properties of the natural world in which the relative speeds of various modes of sensory information differ (Munhall *et al.*, 1996; Gick *et al.*, 2010), a addendum to the previous prediction (3 a) is:

Prediction 3 b: There will be an asymmetry in responses in that increased /pa/ responses to visual-lead stimuli will be sustained over greater SOAs than increased /pa/ responses to tactile-lead stimuli. This is because the speed of light is faster than the speed of airflow during speech.

II. METHODS

Fifty-five University of British Columbia (UBC) students took part in the study and received course credit for their participation. As a result of UBC's policies on participant recruitment for course credit, no restrictions could be placed on native speaker requirements, or whether participants had prior knowledge of the study. Due to the focus of the main task being dependent on a phonological contrast found in English but not necessarily present in other languages, the results from 23 non-native English speakers' data were not analysed in the present study. Of the 32 remaining native English speakers, one was excluded due to experimenter error (forgetting to turn on the babble audio) and five were excluded because they had prior knowledge of the study's purpose. Of the remaining 26 participants, the age range was 18–40 years, $M = 21.23$, $SD = 4.67$ years, with 19 females. Participants gave informed consent and reported no history of speech or hearing issues.

Each participant completed a visual-tactile integration task and a language background questionnaire. The visual-tactile integration task was a two-alternative forced-choice response task administered using PSYCHOPY (Peirce, 2007) experimental software. Participants watched silent videos of a person saying /pa/ or /ba/. While watching the videos, they received gentle puffs of air to their skin in some trials (in all but one condition) and heard English, multi-talker babble. Participants were seated in a sound-attenuated booth, with their heads positioned against a headrest to prevent excessive movement. An air tube which released the puff of air was positioned ~7 cm from the suprasternal notch (front of neck) of each participant. During pre-task instructions, they were told they would feel puffs of air on their skin at some points during the experiment. Participants wore direct sound extreme isolation headphones through which they heard continuous English, multi-talker babble. This was to mask any sound coming from the air tube when the air puff was released and to create a more natural speech environment in which the utterances seemed to be inaudible due to the babble, rather than because they were silent speech.

Participants were instructed to watch the person on the screen speaking and respond via keyboard as to what he had said. The two response key options were the *z* and *slash* keys, which were labelled "pa" and "ba." Response keys were counterbalanced across participants. Participants' responses triggered the next trial to appear automatically on screen. Participants completed four practice trials before the task began and no feedback was given. During the task, ten conditions were presented: a visual-only condition and the following SOA conditions: 0 ms (synchronous), ±50 ms, ±100 ms, ±200 ms, ±300 ms, where "+" means that the visual stimulus precedes the tactile stimulus and "−" means the tactile stimulus precedes the visual. Each condition was presented ten times with both /pa/ and /ba/ visual stimuli for a total of 200 tokens which were completely randomized. The task took less than 15 min to complete.

The presentation and timing of the silent videos and puff stimuli were coordinated via a specially designed switchbox, which caused the release of the air puffs when a 10 kHz, 1 db sinewave of 30 ms duration was detected. This sine wave was added as an audio track to the video file and when the video file was played this audio signal was directed via audio cable to the switchbox. It was therefore inaudible to the participant.

To create the visual stimuli, a 28 year old male native speaker of Vancouver English was instructed to produce eight repetitions of /pa/ and /ba/ in isolation, speaking naturally. The productions were recorded on a JVC camcorder (model GZ-E300AU), 48 kHz stereo PCM audio, 24 frames/s video, and 1280 × 720 pixel resolution. Editing proceeded

in Adobe Premiere ProCC. Five productions of each syllable (/pa/ and /ba/) were chosen based on neutral facial expression, naturalness and consistency of production (as judged by the researcher, a native speaker of Australian English). Each spoken syllable was extracted from the original recording, saved as its own video and trimmed to 1800 ms so that the duration of each video was consistent.

The audio track from each video was extracted for analysis in PRAAT (version 5.4.08; Boersma and Weenink, 2015) and removed from the video files. In preparation for creating the tactile stimulus (air puff), the moment of the vowel onset for each production of /pa/ and /ba/ and the burst only for each /pa/ production were determined in PRAAT. The vowel onset was judged as the onset of the periodic portion of the waveform following the release of the stop. The burst for /pa/ productions was determined as the first spike in the waveform after the initial period of silence. The vowel onset was then subtracted from the burst for each production to determine the average voice onset time (VOT) for the speaker's voiceless bilabial stop ($M = 98.97$ ms, $SD = 3.69$). Average VOT for /ba/ syllables was ~10 ms ($M = 9.83$ ms, $SD = 0.54$ ms), but the VOT for /ba/ was not considered in the creation of the tactile stimuli as the air puff was intended to be modelled on the speaker's aspiration in voiceless stops, simulating as closely as possible the same duration, and timing with respect to the vowel onset (for the synchronous condition).

Tactile stimuli were similar to those used by Gick and Derrick (2009), but with an updated switchbox and updated control software. A Jobmate air compressor set to ~6 psi was connected via a 1/4 in. vinyl tube to a custom-made switchbox which housed a solenoid valve. This equipment was situated outside the sound-attenuated booth in which the tasks took place. A second 1/4 in. vinyl tube ran from the switchbox, through the wall of the sound booth and was attached at the other end to a microphone stand with a flexible head. This end released the puff of air, which was directed towards the participant. The microphone stand was placed to the left of the participant and the end of the vinyl tube was positioned ~7 cm from the suprasternal notch of each participant. An audio cable ran from the computer, which played the visual stimuli to the switchbox. When each stimulus was presented, the sine wave from the video file was detected by the switchbox and triggered the switch, which in turn triggered the solenoid valve to open. This process took ~45 ms and resulted in a gentle puff of air being released towards the suprasternal notch of the participant. These compressor settings produce an impact peak of ~3 cm H_2O, or ~1/3 typical air pressure, from conversational speech. Due to the time it took for the solenoid valve to close again after opening, a 30 ms sine wave caused an air puff of 100 ms duration.

To create the stimuli for the synchronous (0 ms) condition, the sine wave was positioned so that its onset occurred 100 ms prior to the vowel onset (determined from the original speech audio) of each production, both /pa/ and /ba/. It was then shifted another 45 ms to the left to account for the switchbox system latency, therefore in total the sine wave onset occurred 145 ms before the vowel onset in each production, as illustrated below in Fig. 1(b). This resulted in the onset of the air puff occurring 100 ms prior to the original vowel onset and ending at the vowel onset, which simulated the timing of the original period of aspiration in the speaker's aspirated syllables. Due to the difference in VOT between /pa/ and /ba/, this meant that the puff was differentially aligned for the unaspirated syllables as compared to aspirated syllables. For example, when the sine wave was positioned for the 100 ms SOA condition, it was actually closer to being synchronous with the burst for /ba/ productions; the /ba/ bursts occurred ~10 ms before the vowel onset so the difference in alignment between the sine wave onset and /ba/ burst onset was only ~10 ms, whereas in the synchronous condition, the sine wave onset occurred ~90 ms before the burst for /ba/. It should, however, be noted that visually the mouth is not always clearly open at the moment of the burst and so the visual cues may differ slightly from the (absent) audio cues. Nevertheless, it was important to maintain consistency of puff duration and onset position across syllables, therefore the voiceless, aspirated syllable was chosen as the model for simulating aspiration. Appropriate adjustments for the position of the sine wave were made for the various SOA conditions. For example, in the −200 ms condition, the onset of the sine wave was positioned 345 ms prior to the vowel onset [see Fig. 1(a)], whereas the onset of the sine wave for the 200 ms condition was positioned 55 ms following the vowel onset [see Fig. 1(c)].

III. RESULTS

Prediction 1 was that participants would perform at chance levels in the visual-only condition. Figure 2 illustrates the percentage of /pa/ and /ba/ responses in this condition.

Performance at chance levels should show equal rates of /pa/ and /ba/ responses; as can be seen, the rate for /pa/ responses is lower (34%). This is a significant deviation from chance, as shown by a binomial test, $p = <0.001$, 95% CI [0.61, 0.70], thus demonstrating that participants exhibit a /ba/ bias in the visual-only condition.

A. Synchronous condition

The introduction of the tactile stimulus in the synchronous puff condition was predicted to produce an increase in /pa/ responses, as compared with the visual-only condition (prediction 2). Figure 3 illustrates the total percentage of /pa/ responses in the visual-only and synchronous puff (0 ms) conditions. The figure shows that in the visual-only condition, /pa/ responses to the stimuli are 34%, whereas in the synchronous puff condition /pa/ responses increase to 59%.

To investigate whether the percentage of /pa/ responses differs significantly between the visual-only and synchronous conditions, a logistic mixed effects model was fit using the GLMER function (Bates *et al.*, 2014) in R (R Core Team, 2014) with response as the dependent variable, visual stimulus type (/pa/ or /ba/) and puff condition (visual-only, synchronous) and their interaction as fixed effects, a random effect for participant and a by-participant random slope for the interaction of visual stimulus type and puff condition, and a second random slope for each visual token. The model

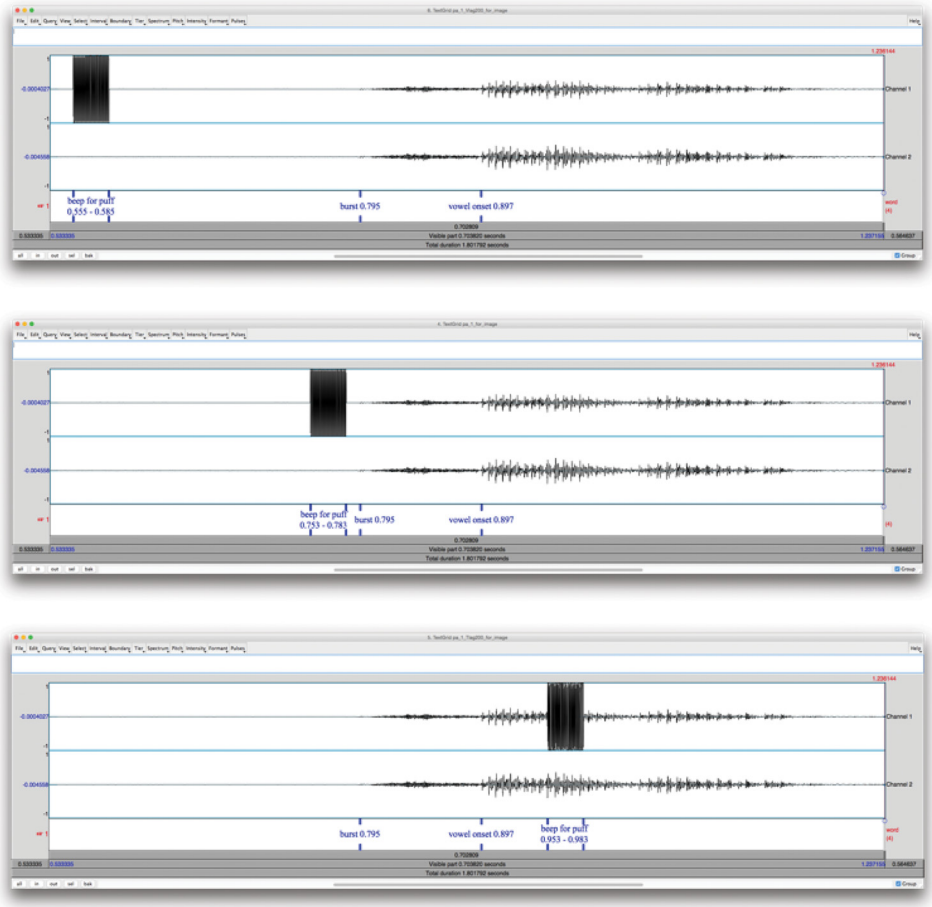

FIG. 1. (Color online) Position of sine wave relative to vowel onset of a /pa/ production (a) −200 ms, (b) 0 ms condition, and (c) 200 ms condition. Note that the audio tracks these images are based on are used only for illustrative purposes. All original speech audio was removed from the experiment stimuli.

formula, as written in R (R core team, 2014), and in the shape of *dependent variable* ∼ *independent fixed-effects* + (*random-effects*|*random variable*) is as follows:

$$\begin{aligned} Response \sim\ & Visual\ stimulus * Puff\ condition \\ & + (1 + (Visual\ stimulus * Puff \\ & condition)|Participant) + 1|Token). \end{aligned}$$

Response refers to the dependent variable (either /ba/ or /pa/). *Visual stimulus* ∗ *Puff condition* refers to the fixed-effects interactions visual stimulus (either video of /ba/ or /pa/) and puff condition (in this case either a puff at 0 ms, or no puff). These are the key variables for the analysis. The first random effect (1 + (*Visual stimulus* ∗ *Puff condition*)|*Participant*) acts in similar ways to a repeated-measures analysis of variance, but without assumptions about normal

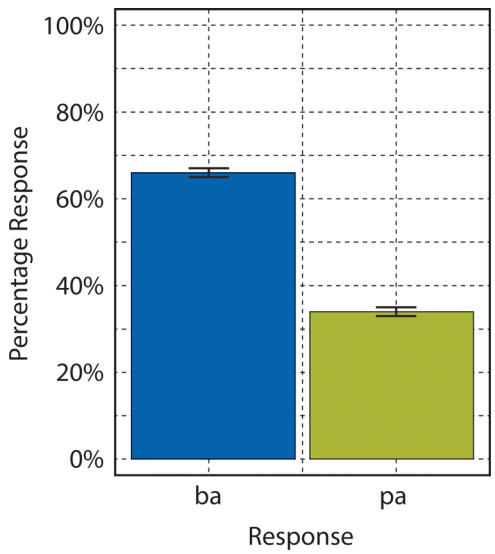

FIG. 2. (Color online) Percentage of /ba/ and /pa/ responses in the visual-only condition.

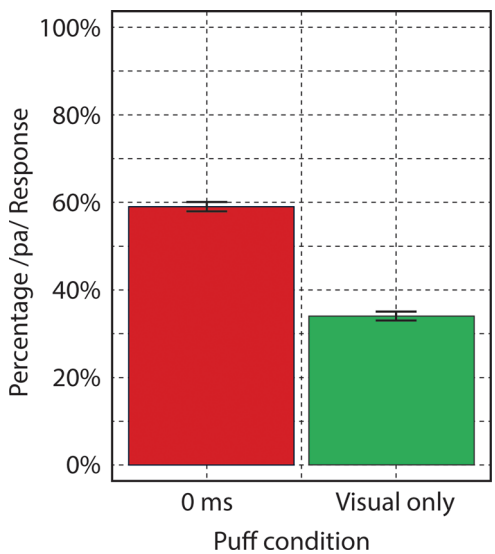

FIG. 3. (Color online) Percentage of /pa/ responses in the 0 ms and visual-only conditions.

distribution or group size. The second random effect $(1|Token)$ factors out idiosyncratic effects of individual video files unrelated to whether they are of /ba/ or /pa/.

Results (see Table I) show a significant main effect of puff condition ($\beta = 0.93$, $SE = 0.28$, $z = 3.32$, $p = <0.001$), indicating that participants report significantly more /pa/ responses during trials where there is a puff of air presented synchronously with the visual stimulus. The interaction between the type of visual stimuli and puff condition also approached significance ($\beta = 0.54$, $SE = 0.28$, $z = 1.92$, $p = 0.054$).

B. Asynchronous conditions and asymmetry

Prediction 3a was that when the SOA was smaller (closer to synchronous), participants would report more /pa/ responses. Related to this, prediction 3b was that the /pa/ responses would be asymmetric; that is, participants would be more likely to perceive a token as /pa/ when the visual stimulus preceded the audio, than when the opposite order occurred. Figure 4 shows the locally weighted scatterplot mooting (LOESS) mean and standard errors (R Core Team, 2014) of /pa/ responses as a function of the SOA and visual stimulus. As illustrated, when the tactile stimulus leads the visual by 300 ms (−300 ms SOA) /pa/ responses are around 40% for both /pa/ and /ba/ visual stimuli, with /pa/ visual stimuli having slightly higher percentage /pa/ responses (42%) as compared to /ba/ visual stimuli (37%). As the SOA progresses towards synchronous (0 ms SOA), there is a rise in /pa/ responses which reaches around 60% at 0 ms SOA (64% when the visual stimulus is /pa/ and 55% when the visual stimulus is /ba/). As can be seen, there is a clear asymmetry in responses. Figure 4 also shows that the highest rate of /pa/ responses does not occur when the two stimuli are synchronously presented. When the visual stimulus is /pa/, the highest /pa/ response (65%) is at 50 ms SOA and when the visual stimulus is /ba/ the highest /pa/ response (67%) is further rightwards, at 200 ms SOA. The rate of /pa/ responses can be seen to drop off after the respective peaks.

To investigate whether the degree of asynchrony of the SOAs significantly affects responses and whether this happens in a symmetrical fashion, a logistic mixed effect model was fit with response as the dependent variable, visual stimulus and SOA (0 ms, ±50 ms, ±100 ms, ±200 ms, ±300 ms) and their interaction as fixed effects (SOAs were converted from factors to a scale as the SOAs were varied along a continuum and were thus related), a random effect for participant and a by-participant random slope for the interaction of visual stimulus and SOA, with a second random slope accounting for each visual token. The model formula, as written in R, is as follows:

$$Response \sim Visual\ stimulus * SOA + (1 + (Visual\ stimulus * SOA)|Participant) + (1|Token).$$

This model (see Table II) shows a significant main effect of SOA ($\beta = 0.43$, $SE = 0.10$, $z = 4.43$, $p = <0.001$), indicating that as the SOA increases, there is also an increase in /pa/ responses. There is also an interaction between the visual stimulus and SOA ($\beta = -0.23$, $SE = 0.09$, $z = -2.58$, $p = 0.009$) such that if the stimuli is a /pa/, the peak integration is closer to 0 ms, as visualized in Fig. 4 above.

A polynomial model, as written in R, with the structure described below was run to investigate whether the various SOAs significantly affected responses in symmetrical fashion,

$$Response \sim Visual\ stimulus * Poly(SOA, degree = 2) + (1|Participant) + (1|Token).$$

Results are shown in Table III below.

The results show that like the linear model, the second order polynomial model demonstrates a significant main effect of the intercept and SOA. In addition, there was also a significant interaction between visual stimuli type and SOA.

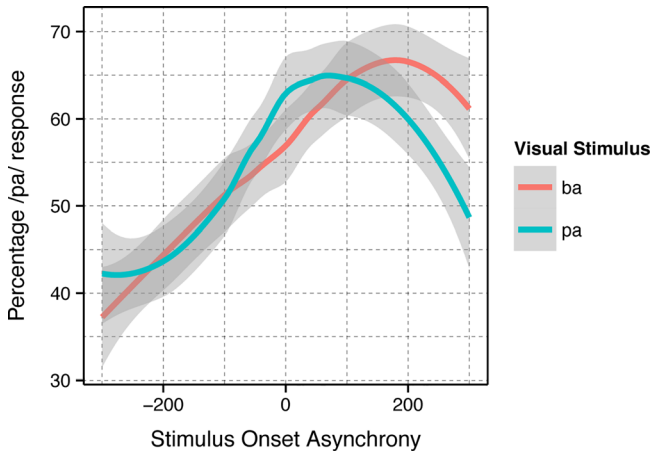

FIG. 4. (Color online) Percentage of /pa/ responses across SOAs, split by visual stimulus.

TABLE I. Results of mixed model with structure: $Response \sim Visual\ stimulus * Puff\ condition + (1 + (Visual\ stimulus * Puff\ condition)|Participant) + (1|Token)$.

	β	SE	z	p
Intercept	−0.71	0.28	−2.55	0.011[a]
Visual stimulus (pa)	−0.08	0.35	−0.24	0.808
Puff condition (0 ms)	0.93	0.28	3.32	<0.001[b]
Visual stimulus (pa): Puff condition (0 ms)	0.54	0.28	1.92	0.054

[a] $p < 0.1$.
[b] $p < 0.001$.

TABLE II. Results of mixed model with structure: $Response \sim Visual\ stimulus * SOA + (1 + Visual\ stimulus * SOA|Participant) + (1|Token)$.

	β	SE	z	p
Intercept	0.27	0.23	1.15	0.25
Visual stimulus (pa)	−0.05	0.32	−0.16	0.87
SOA	0.43	0.10	4.43	<0.001[a]
Visual stimulus (pa): SOA	−0.23	0.09	−2.58	0.009[b]

[a] $p < 0.001$.
[b] $p < 0.01$.

To test whether responses at each individual SOA differ significantly from the visual-only condition a model was run with fixed effects for SOA, visual stimulus and their interaction, as well as a random variable for participant and token. SOAs were converted to factors as this allowed for significance testing at each SOA. That model failed to converge. The model was simplified until the only remaining random variable was the one for token, whereupon it converged. The model formula, as written in R, is as follows:

$$Response \sim (Visual\,Stimulus * SOA) + (1|Token).$$

Results (see Table IV) show that the temporal window of visual-tactile integration is asymmetric, with significantly higher /pa/ responses from −200 ms to 300 ms, compared to the visual-only condition.

IV. DISCUSSION

This study examined the potential integration of aero-tactile speech information during visual-tactile speech perception. Based on previous research into audio-visual and audio-tactile speech perception, several predictions were made with respect to participants' perceptual behaviour when presented with visual-tactile speech. Prediction 1 held that participants would perform at chance levels when presented with visual-only bilabial syllables, but found that when presented with visual-only /pa/ or /ba/, they were significantly more likely to report the unaspirated syllable /ba/, counter to the prediction. This might be a somewhat unexpected result as the speech sounds /b/ and /p/ are generally considered to be instances of a single viseme (Fisher, 1968). In addition, a similar perception study (Abel *et al.*, 2011) found a response bias against /b/ (though a three-way versus two-way identification task may not be a legitimate comparison). However, as discussed in the introduction, the result fits the pattern seen in all of the previous research on audio-(aero)tactile integration (Gick and Derrick, 2009; Gick *et al.*, 2010; Derrick and Gick, 2013). There is a bias towards /ba/ in two-way forced-choice experiments with /pa/ vs /ba/ for both unimodal auditory and visual perception. This result of a /ba/ bias in the visual-only condition is the baseline for comparison with all of the other conditions containing the tactile stimulus, somewhat limiting the statistical power of the subsequent analyses.

TABLE III. Results of mixed model with structure: *Response* ∼ *Visual stimulus* ∗ *Poly*(*SOA, degree* = 2) + (1|*Participant*) + (1|*Token*).

	β	*SE*	*z*	*p*
Intercept	0.24	0.23	1.04	0.30
Visual stimulus (pa)	−0.03	0.31	−0.10	0.92
Poly(SOA, degree = 2)1	26.21	3.14	8.35	<0.001[a]
Poly(SOA, degree = 2)2	−10.15	3.17	−3.20	0.001[b]
Visual stimulus (pa): Poly(SOA, degree = 2)1	−13.02	4.35	−3.03	0.002[b]
Visual stimulus (pa): Poly(SOA, degree = 2)2	−7.51	4.42	−1.67	0.089

[a] $p < 0.001$.
[b] $p < 0.01$.

TABLE IV. Results of mixed model with structure: *Response* ∼ (*Visual Stimulus* ∗ *SOA*) + (1|*Token*).

	β	*SE*	*z*	*p*
Intercept	−0.65	0.25	−2.66	<0.008[a]
Visual stimulus (pa)	−0.06	0.35	−0.18	0.85
SOA −300 ms	0.08	0.19	0.47	0.63
SOA −200 ms	0.40	0.19	2.14	0.032[b]
SOA −100 ms	0.67	0.19	3.58	<0.001[c]
SOA −50 ms	0.92	0.19	4.91	<0.001[c]
SOA 0 ms	0.87	0.19	5.64	<0.001[c]
SOA 50 ms	1.14	0.19	6.05	<0.001[c]
SOA 100 ms	1.35	0.19	7.09	<0.001[c]
SOA 200 ms	1.39	0.19	7.26	<0.001[c]
SOA 300 ms	1.142	0.19	6.05	<0.001[c]
Visual stimulus (pa): SOA −300 ms	0.31	0.26	1.19	0.23
Visual stimulus (pa): SOA −200 ms	0.05	0.26	0.19	0.84
Visual stimulus (pa): SOA −100 ms	0.07	0.26	0.26	0.79
Visual stimulus (pa): SOA −50 ms	0.02	0.26	0.08	0.93
Visual stimulus (pa): SOA 0 ms	0.45	0.26	1.70	0.09
Visual stimulus (pa): SOA 50 ms	0.22	0.27	0.85	0.40
Visual stimulus (pa): SOA 100 ms	−0.09	0.27	−0.35	0.73
Visual stimulus (pa): SOA 200 ms	−0.23	0.27	−0.86	0.39
Visual stimulus (pa): SOA 300 ms	−0.26	0.26	−1.85	0.06

[a] $p < 0.01$.
[b] $p < 0.1$.
[c] $p < 0.001$.

A major purpose of this study was to examine whether people integrate aero-tactile information during visual-tactile speech perception in such a way that it affects their categorization. A similar finding has been made for audio-visual (McGurk and MacDonald, 1976) and audio-tactile (Gick and Derrick, 2009) modality combinations. The results comparing the visual-only and synchronous conditions supported the major prediction (2). The results demonstrate that the puff of air significantly affected the perception of the visual stimulus so that in the synchronous puff condition, participants were more likely to perceive a voiceless stop /pa/ (phonetically a voiceless aspirated stop), as opposed to a voiced stop /ba/ (phonetically a voiceless unaspirated stop), than in the conditions where the visual stimulus was presented in isolation. As noted above, however, one limitation of this finding was that participants did not perform at chance levels in the visual-only condition. Nevertheless, these findings appear to show that people integrate this aero-tactile information as the sensation of aspiration. This suggests that, whether speakers are consciously aware of it or not, the aero-tactile mode provides a portion of the information that constitutes the speech stream and that this information is integrated with that received through other modalities to contribute to the overall perception of the speech signal. This result also shows that speech integration can occur without participants receiving any information at all from the audio speech signal. The findings from this study also show that combinations of visual and aero-tactile speech information contribute enough to the speech stream to affect perceivers' categorization of speech segments. This is an informative result considering the audio signal is generally considered to be the primary source of speech information.

However, it is in line with the results of Sumby and Pollack (1954) results that show if the audio signal-to-noise ratio is lower than about −12 db, the visual component contributes more to accurate speech perception than the audio component.

This study also examined whether the degree of asynchrony of the stimuli affected participants' responses. Based on previous research (Munhall *et al.*, 1996; Gick *et al.*, 2010), it was predicted that the closer the SOA was to synchronous, the more likely participants would be to consider the visual and tactile stimulus combination as being part of the same event, and therefore the more likely they would be to integrate the tactile stimulus and report a /pa/ (prediction 3a). Figure 4 showed that at −300 ms where the tactile stimulus leads the visual stimulus, participants reported perceiving /pa/ at around 40% and this rate increased steadily as the SOA grew closer to synchronous, peaking at 50 ms when the visual stimulus was /pa/ and 200 ms when the visual stimulus was /ba/. The /pa/ responses declined again towards 300 ms, though they were noticeably higher at 300 ms than −300 ms. Participant responses at SOAs of −200 ms through to 300 ms showed a significant difference from responses in the visual-only condition, with the strength of the significance peaking at 100–200 ms. These findings showed that although /pa/ responses increased towards the synchronous condition, they peaked later, in partial validation of prediction 3a.

These findings also suggest a group window of visual-tactile integration of −200 to 300 ms, considerably wider than that found for audio-visual stimuli (0 to 180 ms for Munhall *et al.*, 1996 and −30 to 170 ms for van Wassenhove *et al.*, 2007) and audio-tactile stimuli (−50 to 200 ms for Gick *et al.*, 2010). The results from all these studies share two qualities: The first is they are same in terms of the direction of asymmetry, matching with what can happen in nature when a speaker is far enough away from the perceiver. The second is that windows of integration almost always extend beyond the boundaries of what can occur during natural speech production. We believe this is a by-product of signal ambiguity, and would be more pronounced the more ambiguous the combined signals. However, the cause is not yet known, and is a good object for future research.

The results also show that the window of integration, based on perceptions of the speaker in the present study, is significantly narrower for visual /pa/ stimuli. The difference in responses when the visual stimulus was /pa/ compared to /ba/ may have been due in part to the position and duration of the tactile stimulus, which as previously discussed was designed to be aligned with /pa/ visual stimuli, but not /ba/ stimuli, which has a shorter VOT. However, this result contradicts Fisher's (1968) notion that labial /p/ and /b/ are instances of a single viseme, supporting the part of the research of Abel *et al.* (2011) that shows a visual distinction between the 2. More generally, the result also fits in with previous research showing people are not limited to perceiving viseme categories (Bernstein and Liebenthal, 2014; Auer and Bernstein, 1997). Those previous results show that there is subtle, dynamic information contained within visual speech that can help with accurate speech reading, and our results show that this information can influence the window of perceptual integration.

Perceivers do identify subtle cues within the visual signal; perceptual integration of visual /pa/ and aspiration was narrower than perceptual interference of visual /ba/ and aspiration. This result is the opposite of that obtained for audio-(aero)tactile integration (Gick *et al.*, 2010), where the window of integration was longer than the window of interference. Note that this may be an appropriate pattern if one considers that for this experiment, the auditory stream contains more information than the aero-tactile stream, which contains more information than the visual stream. However, the cause is not yet known, and worthy of future research.

Prediction 3b was that participants would integrate speech information from the visual and tactile modalities over a greater range of SOAs when the visual stimulus preceded the tactile stimulus, as opposed to when the stimuli were presented in the opposite order. This prediction was based on previous research, which found asymmetry in the windows of multimodal integration (Munhall *et al.*, 1996; Gick *et al.*, 2010), as well as on knowledge of physical properties of the world such as speeds of light and speed of speech airflow. Results showed that in support of this prediction, participants were more likely to report the voiceless aspirated stop when presented with visual-leading tokens. These results were seen in Fig. 3 where the rate of /pa/ responses (to both /pa/ and /ba/ visual stimuli) was generally higher when the visual stimulus preceded the tactile stimulus. This suggests that, similarly to perception of audio-visual and audio-tactile speech stimuli, individuals perceive visual-tactile speech stimuli in a manner consistent with relative speeds of physical properties of the world. The range of integration on the right side was noticeably wider than was found for audio-visual and audio-tactile combinations, an observation that may be attributed to the very large difference in transmission speed of speech airflow as compared to the speed of light.

The perceptual behaviour observed for the asynchronous stimuli was apparent even though speakers could not have had experience perceiving aspiration from such a great distance, as measurable airflow from aspiration is known to dissipate by around 30–40 cm from the mouth (Derrick *et al.*, 2009). Aspiration airflow is known to be delayed by 25 ms at 17 cm and by 100 ms at between 30 and 35 cm distance (Derrick *et al.*, 2009). Thus, while participants may have had some experience perceiving slightly delayed aspiration with respect to the rest of the speech signal at 50 ms, and possibly even at 100 ms, two of the SOAs tested, their responses at 200 and 300 cm distance cannot be based on direct experience of speech. Perceivers may, however, have had analogous non-speech experiences that could underscore delays in perceiving aero-tactile stimuli as compared to visual stimuli of the same event, e.g., feeling the delayed air from an oscillating fan across a room when the fan has already turned in another direction. This more general airflow information may contribute to speakers' knowledge of how airflow from aspiration behaves. Alternatively, the present results may be seen as adding a sufficient level of complexity to bring into question a learned-mapping approach to multisensory perception. It may instead favor a model that more naturally engenders rich multimodality in perception (e.g., Fowler, 1986).

Regarding the question of prior speaker knowledge of speech aspiration, speakers arguably do have considerable understanding of aspiration behaviour. Several communicative situations provide direct tactile stimuli in the form of aspiration. Whispering is one example; in this situation the listener may receive aspiration information to the ear or the side of the face. Simply speaking normally in close proximity to a conversation partner may also be a situation where aspiration can be felt. Other situations provide a visual representation of aero-tactile speech information which can be incorporated as general knowledge of aspiration, as was shown by Mayer *et al.* (2013) with the visible perturbation of a candle. Further examples include speaking in very cold temperatures, where speakers produce visible puffs of air, and speaking while smoking. The use of microphones is a situation in which aspiration is represented audibly. All of these experiences may contribute to speakers' awareness of the behaviour of aspiration. Speakers may also receive aero-tactile feedback to their own lips when they speak. Results of the current study suggest that naive speakers can identify this aspiration information with particular speech segments in their language.

Findings from the present study have shown that perceivers have either conscious or unconscious awareness of aero-tactile speech information, an underexplored area in speech research. Perceivers have shown the ability, in the absence of an audible speech signal, to make use of information from the aero-tactile mode to distinguish speech sounds when they are presented with an ambiguous visual speech signal. This ability shows that aero-tactile speech information is influential in multimodal speech integration and contributes enough information to the signal to shift categorization of speech segments. It also demonstrates that multimodal speech integration occurs even without an audio signal. The novel pairing of modalities corroborates the prediction that signal speed is a determining factor in asymmetric windows of integration. The current findings also suggest that speech may be better characterized as being perceived modality neutrally (though the weighting of speech-relevant information may vary by perceiver or by sounds).

ACKNOWLEDGMENTS

This research was funded by a Discovery Grant from the Natural Sciences and Engineering Council of Canada to B.G., and by National Institutes of Health Grant No. DC-02717 to Haskins Laboratories. Thank you to Eric Vatikiotis-Bateson, Kathleen Currie-Hall, Molly Babel, Martina Wiltschko, Nicole Anger, Ryan Hill, Michael McAuliffe, Avery Ozburn, Megan Keough, Bosko Radanov, and Oksana Tkachman, for advice and support throughout this research. Special thanks to Nick Romero for making the air pump switchbox, and also to everyone in the Interdisciplinary Speech Research Lab at UBC.

Abel, J., Barbosa, A. V., Mayer, C., and Vatikiotis-Bateson, E. (**2011**). "The labial visemereconsidered: Evidencce from production and perception," in *9th International Seminar on Speech Production (ISSP)*, edited by Y. Laprie and I. Steiner, Montreal, Quebec, pp. 337–344.

Alcorn, S. (**1932**). "The Tadoma method," Volta Rev. **34**, 195–198.

Auer, E. T., Jr., and Bernstein, L. E. (**1997**). "Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness," J. Acoust. Soc. Am. **102**, 3704–3710.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (**2014**). "From _lme4: Linear mixed-effects models using Eigen and S4_," R package version 1.1-7, http://CRAN.R-project.org/package=lme4 (Last viewed 2/13/2016).

Bernstein, L. E., and Liebenthal, E. (**2014**). "Neural pathways for visual speech perception," Front. Neurosci. **8**(386), 1–18.

Boersma, P., and Weenink, D. (**2015**). "Praat: Doing phonetics by computer" [computer program], version 5.4.08, http://www.praat.org/ (Last viewed 3/24/2015).

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D., and David, A. S. (**1997**). "Activation of auditory cortex during silent lipreading," Science **276**(5312), 593–596.

Derrick, D., Anderson, P., Gick, B., and Green, S. (**2009**). "Characteristics of air puffs produced in English "pa": Experiments and simulations," J. Acoust. Soc. Am. **125**(4), 2272–2281.

Derrick, D., and Gick, B. (**2013**). "Aerotactile integration from distal skin stimuli," Multisens. Res. **26**, 405–416.

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

Fisher, C. (**1968**). "Confusion among visually perceived consonants," J. Speech Hear. Res. **11**, 796–804.

Fowler, C., and Dekle, D. (**1991**). "Listening with eye and hand: Cross-modal contributions to speech perception," J. Exp. Psychol.: Hum. Percept. Perform. **17**(3), 816–828.

Fowler, C. A. (**1986**). "An event approach to the study of speech perception from a direct-realist perspective," in *Status Report on Speech Research*, edited by I. G. Mattingly and N. O'Brien, Haskins Laboratories, New Haven, CT, pp. 139–169.

Fowler, C. A. (**2004**). "Speech as a supramodal or amodal phenomenon," in *The Handbook of Multisensory Processes*, edited by G. A. Calvert, C. Spence, and B. E. Stein (MIT Press, Cambridge, MA), pp. 189–201.

Gick, B., and Derrick, D. (**2009**). "Aero-tactile integration in speech perception," Nature **462**(7272), 502–504.

Gick, B., Ikegami, Y., and Derrick, D. (**2010**). "The temporal window of audio-tactile integration in speech perception," J. Acoust. Soc. Am. **128**(5), EL342–EL346.

Gick, B., Jóhannsdóttir, K., Gibraiel, D., and Mühlbauer, J. (**2008**). "Tactile enhancement of auditory and visual speech perception in untrained perceivers," J. Acoust. Soc. Am. **123**(4), EL72–EL76.

Hayden, R. E. (**1950**). "The relative frequency of phonemes in general-American English," Word **6**(3), 217–223.

Mayer, C., Gick, B., Weigel, T., and Whalen, D. (**2013**). "Perceptual integration of visual evidence of the airstream from aspirated stops," Can. Acoust. **41**(3), 23–27.

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

Munhall, K., Gribble, P., Sacco, L., and Ward, M. (**1996**). "Temporal constraints on the McGurk effect," Percept. Psychophys. **58**(3), 351–362.

Peirce, J. (**2007**). "PsychoPy—Psychophysics software in Python," J. Neurosci. Meth. **162**(1-2), 8–13.

R Core Team. (**2014**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/ (Last viewed 2/13/2016).

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

Rosenblum, L. D., Pisoni, D. B., and Remez, R. (**2005**). "Primacy of multimodal speech perception," in *The Handbook of Speech Perception* (Blackwell, Malden, MA), pp. 51–78.

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

van Wassenhove, V., Grant, K., and Poeppel, D. (**2007**). "Temporal window of integration in auditory-visual speech perception," Neuropsychologia **45**(4), 598–607.