The effect of lip-reading on primary stream segregation

Aymeric Devergie and Nicolas Grimault[a)]
Centre de Recherche en Neurosciences de Lyon, UMR CNRS 5292 Université Lyon 1, 69366 Lyon Cedex 07, France

Etienne Gaudrain
MRC Cognition and Brain Sciences Unit, Cambridge, United Kingdom

Eric W. Healy
Speech Psychoacoustics Laboratory, Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210-1002

Frédéric Berthommier
GIPSA-Lab CNRS UMR 5216, Domaine universitaire 38402 Saint Martin d'Hères, France

(Received 7 June 2010; revised 16 March 2011; accepted 25 April 2011)

Lip-reading has been shown to improve the intelligibility of speech in multitalker situations, where auditory stream segregation naturally takes place. This study investigated whether the benefit of lip-reading is a result of a primary audiovisual interaction that enhances the obligatory streaming mechanism. Two behavioral experiments were conducted involving sequences of French vowels that alternated in fundamental frequency. In Experiment 1, subjects attempted to identify the order of items in a sequence. In Experiment 2, subjects attempted to detect a disruption to temporal isochrony across alternate items. Both tasks are disrupted by streaming, thus providing a measure of primary or obligatory streaming. Visual lip gestures articulating alternate vowels were synchronized with the auditory sequence. Overall, the results were consistent with the hypothesis that visual lip gestures enhance segregation by affecting primary auditory streaming. Moreover, increases in the naturalness of visual lip gestures and auditory vowels, and corresponding increases in audiovisual congruence may potentially lead to increases in the effect of visual lip gestures on streaming.
 [DOI: 10.1121/1.3592223]

PACS number(s): 43.66.Mk, 43.71.Rt [KWG] Pages: 283–291

I. INTRODUCTION

Previous research has explored the segregation mechanisms that are most likely to be employed in competitive listening situations, such as the perception of concurrent speech (Bregman, 1990). van Noorden (1975) described a helpful experimental paradigm to study the contribution of acoustic cues to auditory segregation and, specifically, sequential segregation mechanisms. This streaming paradigm uses the sound sequence, ABA-ABA-···, composed of two tones, A and B, that differ by some acoustic attribute. Moore and Gockel (2002) found that any salient acoustic difference can help listeners to segregate A tones from B tones and to group them into two distinct auditory streams. In his work, van Noorden (1975) observed two types of streaming mechanisms depending on the task given to the subject. Obligatory, automatic or primary streaming is observed when subjects try to fuse the sequence into a single stream (but fail to do so), whereas voluntary or schema-based streaming is observed when subjects try to segregate the sequence into two streams (and succeed in doing so). In addition, using a similar paradigm, Bregman (1978) reported that segregation requires about two or three seconds of build-up time to take place.

More recently, electrophysiological studies have been conducted to determine the level of processing at which primary stream segregation takes place. The method used in these studies was to record neural firing rate during the presentation of ABA sequences. Because segregation was initially absent due to build-up, the ABA sequence paradigm enabled researchers to compare different segregation states (i.e., items integrated versus segregated). Using this method, Micheyl *et al.* (2005) recorded single unit responses in the primary auditory cortex of awake rhesus monkeys. Pressnitzer *et al.* (2008) employed the same recording method in the cochlear nucleus of anesthetized guinea pigs. In these two studies, the authors reported two different firing-rate patterns before and after the build-up period: the units responded to all the A and B tones at the beginning of the sequence, but responded selectively to the A tones after 10 s of build-up. These studies therefore indicate that segregation can take place in the primary sub-cortical and cortical structures of the auditory pathway.

In natural environments, a fundamental contribution to the perception of concurrent speech is lip-reading (speech reading). Lip-reading can improve the intelligibility of speech presented in a noisy environment by up to 40% (Sumby and Pollack, 1954). This benefit is likely sustained by multiple levels of interaction in the integration of audiovisual speech. Massaro and Cohen (1983) and Brancazio and Brancazio (2004) found evidence for high level interactions.

[a)]Author to whom correspondence should be addressed. Electronic mail: nicolas.grimault@olfac.univ-lyon1.fr

Additionally, behavioral and neurophysiological studies have suggested that audiovisual interactions can also occur at lower levels of processing. Along with previous studies (Grant and Walden, 1996; Grant and Seitz, 2000; Grant, 2001; Grant *et al.*, 2004), Bernstein *et al.* (2004) found that speech detection in a noisy environment could be enhanced by visual cues that were synchronized with the sound intervals. Moreover, this effect was larger for lip-reading cues that were highly congruent with the auditory input than for other less congruent visual displays.

Bernstein *et al.* suggested that the benefit reported in a two-interval forced choice detection task could rely on an audiovisual interaction that might have occurred at a relatively primary level of processing. Using fMRI or EEG, several studies have also reported that the presentation of visual articulatory gestures with sounds can activate primary auditory cortical structures (e.g., Pekkola *et al.*, 2005; Kayser *et al.*, 2008; Besle *et al.*, 2008; van Wassenhove *et al.*, 2005). These studies suggest that visual and auditory inputs might interact in primitive neural structures. Since these primary structures seem to support both the primary segregation mechanism and the audiovisual interactions, it can be hypothesized that primary auditory segregation could be modulated by audiovisual interactions.

To further explore the mechanisms of segregation, Rahne and his colleagues (Rahne *et al.*, 2007, 2008) and Rahne and Böckmann-Barthel (2009) built sequences of pure tones designed to induce different perceptual organizations. The frequency of the tones alternated between low and high. Whereas the high-frequency tones appeared in random order, the low-frequency tones together formed a sequence composed of a repeated pattern of three tones rising in pitch, sometime replaced by a deviant pattern of three tones falling in pitch. In addition, every third tone in the overall sequence was more intense by 15 dB. Perceptual organization could then be based on either a frequency difference (grouping the lower tones into one stream and the higher tones into another stream) or an intensity difference (grouping the louder tones into one stream and the softer tones into another stream). A visual cue (squares or circles of different sizes) synchronized either with the frequency or with the intensity pattern was added to influence perceptual organization. In one condition, the visual cues promote segregation based on frequency, signaling the deviant pattern whatever the intensity variations. In the other condition, the visual cue promoted segregation based on intensity whatever the frequency variation.

In two electroencephalography studies, Rahne *et al.* (2007) and Rahne and Böckmann-Barthel (2009) reported mismatch negativity (MMN) when the visual cue promoted a perceptual organization based on a pitch difference, indicating that the participants were indeed perceiving the low-frequency tone sequence segregated from the high-frequency tone sequence, and the resulting deviant pattern. They concluded that a visual cue can alter the perceptual organization of an ambiguous tone sequence. However, as acknowledged by the authors, this design allowed detection of only the pitch based segregated percept, which also correspond to the intensity based integrated percept. It is then unclear from this result whether a visual cue can promote integration across an intensity difference, a pitch difference, or both. Rahne *et al.* (2008) extended these results using the same materials and a behavioral design. Participants were instructed to attend to the visual stimuli and to indicate the currently prevailing sound organization (grouped based on frequency or grouped based on intensity) by pressing one of two buttons on a keypad and to change buttons if the organization changed. The time during which the lower and the higher tones were grouped together was then measured. Their results indicated that the visual cue aimed to promote segregation on the basis of pitch did not affect the perceptual organization. In contrast, the visual cue synchronized with intensity variations succeeded in reducing frequency-based segregation in experimental conditions with a large frequency difference. So, although these authors demonstrated a clear influence of an arbitrary visual cue on the perceptual organization of pure tone sequences, it remains unclear how this influence operates, and it is difficult to extend these conclusions to auditory processing in more ecological situations.

In the current study, sequences of French vowels, as a first approximation of continuous speech, with alternating fundamental frequency (F0) were presented to listeners who perceived either a single stream or two streams, depending on the F0 difference between alternate vowels. In the first experiment, participants had to recall the order of the vowels presented in the sequence (order-naming task). In the second experiment, participants had to detect a change in the rhythm of presentation of the sequence (isochrony detection task). The participants can only succeed in these two tasks if they can integrate the vowels into a single auditory stream (Micheyl and Oxenham, 2010). Therefore, poor performance in these tasks indicates that streaming has occurred despite the effort of the participant to prevent it, thus providing a measure of obligatory (i.e., primary) streaming.

To evaluate the contribution of lip-reading to primary streaming, lip gestures articulating alternate vowels of the sequence were presented to participants in both experiments. We hypothesized that performance would be degraded if the visual and auditory inputs interacted at a low level of processing, thus indicating that primary auditory segregation is enhanced by lip-reading.

II. EXPERIMENT 1

Experiment 1 was designed to test for an effect of lip-reading on primary auditory segregation. Segregation was estimated by assessing participants ability to correctly report the order of presentation of sequences of vowels alternating in pitch. Good performance on this task would suggest that participants integrated all the vowels of the sequence into a single stream (Gaudrain *et al.*, 2007, 2008). To test for a lip-reading effect, visual lip gestures articulating alternate vowels were simultaneously displayed with three different degrees of audiovisual congruence (C_0, C_1, and C_2). In the first condition (C_0), the lips were displayed throughout the vowel sequence without moving. Thus, there was no audiovisual congruence. In the second condition (C_1), the visual lip gestures provided only rhythmic information by opening and closing the mouth in synchrony with the auditory

vowels, causing the audiovisual congruence to be limited solely to temporal aspects. In the last condition (C_2), auditory vowels and visual lip gestures were rhythmically and phonetically congruent. If lip-reading can promote primary segregation, the presence of a congruent visual cue should result in poorer performance.

A. Participants

Ten participants aged between 18 and 24 yr (mean = 20.8, SD = 1.8) took part in the experiment. All of the participants were native French speakers and had pure tone audiometric thresholds below 15 dB hearing level (HL) at octave frequencies between 250 and 4000 Hz (American National Standards Institute, 2004). Participants signed an informed consent form and were reimbursed for their time. This study was formally approved by a local ethics committee (CPP Sud-Est II No. 06035).

B. Stimuli

Sequences of six French vowels (/a/, /e/, /i/, /o/, /y/, /u/) with alternating high and low fundamental frequencies were constructed (Fig. 1). The lowest fundamental frequency F0(1) was equal to 100 Hz. The alternate fundamental frequencies F0(2) were set to one of ten values between 100 and 238 Hz. Each vowel was 166 ms long, including a 10 ms raised-cosine onset and offset ramp, and was adjusted to an RMS value of 85 dB sound pressure level (SPL). These vowels were generated using the Klatt algorithm (Klatt, 1980) and had the same formant values used by Gaudrain *et al.* (2007). A sequence of visual lip gestures pronouncing alternate vowels was presented simultaneously with the audio sequence. These visual lip gestures were synthesized using video-recorded frames as in Berthommier (2003). One hundred video-recorded frames were entered into an XY diagram. The X dimension reflected the horizontal extension of the lips and Y reflected the vertical extension. The starting point (closed lips) and ending point (target position of the lips articulating one particular vowel) were selected manually.

An algorithm by Berthommier (2003) was used to estimate the trajectory between these two points and select the frames closest to it. Thus, the video frames displaying the lips were real but the trajectories were artificial. The decision was made to use these established synthetic visual vowels (after Berthommier, 2003) to test the effect of lip-reading on primary streaming. Auditory and visual dimensions of the stimuli were built separately and then combined.

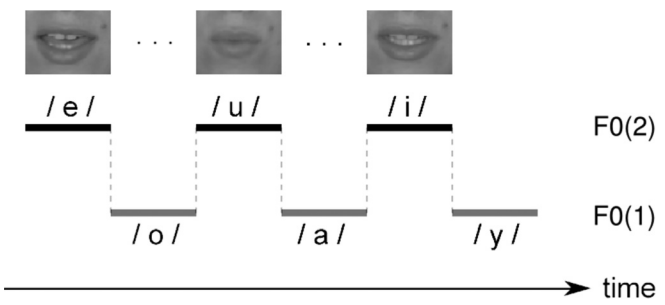

FIG. 1. Schematic representation of an audiovisual sequence. Lip gestures were presented that articulated either the three high-pitch vowels or the three low- pitch vowels, selected randomly across trials [except that in some cases, F0(1) = F0(2)].

Three different types of lip gestures were presented. The first condition (no congruence, C_0) consisted of lips that remained closed during the whole sequence. The second condition (temporal congruence, C_1) consisted of an open-close lip gesture. The lip gesture for the /a/ vowel, which is a neutral open-close gesture, was used for alternate vowels. This visual condition provided a rhythmic cue to either the F0(1) or F0(2) vowels but no phonetic cue. The last condition (temporal and phonetic congruence, C_2) consisted of lip gestures pronouncing the particular F0(1) or F0(2) auditory vowel. In addition to the rhythmic cue provided in C_1, C_2 also provided some phonetic information about the alternate vowel.

The six different lip gestures corresponding to the six different vowels are plotted in the Appendix A (Fig. 6). The relative timing of the audio and video material is detailed in Fig. 2. The lip gestures started 67 ms (i.e., 2 frames at a rate of 30 frames per second) before the corresponding audio vowel. The closing gesture of the lips occurred during the auditory vowel that immediately followed (see Fig. 2). The choice of this temporal offset was made arbitrarily, but was sufficient to preserve a good subjective synchronization between the auditory and visual dimensions of the signals. Also, it has been well established that delays of up to 170 ms for the auditory signal relative to the video can be employed without affecting the binding of the two signals (Grant *et al.*, 2003).

All sequences were created and stored using a C program prior to conducting the experiments. The stimuli were played diotically using a Digigram VxPocket440 sound card connected to a Sennheiser HD 250 Linear II headphone. Visual stimuli were displayed on a monitor with a visual angle of approximately 6°. The listeners were comfortably seated in a double-walled, sound attenuated booth. The output level was calibrated to 85 dB SPL [Larson Davis AEC101 and 824; American National Standards Institute (1995)].

C. Procedure

The participants began the experiment with an identification test to ensure that the vowels were correctly identified. Each vowel sound was played separately, and the participants selected the corresponding vowel among six choices displayed on a computer screen using orthographic representations ("a, é, i, o, u, ou"). Each of the six vowels was played five times, at random F0s among 100, 147, and 238 Hz. All the participants correctly identified the vowels 100% of the time, with the exception of one participant who correctly identified the vowels 87.5% of the time. Next, the participants engaged in the order naming task in which they were briefly trained before the start of testing. To free the participants from having to memorize the sequence, each audiovisual sequence was loop repeated for 10 s. Two seconds after the beginning of each sequence, which is

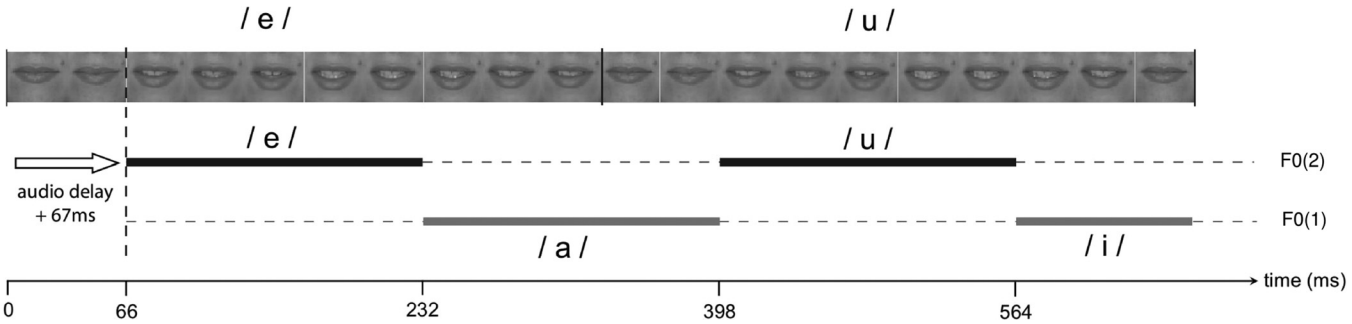

FIG. 2. Synchronization of audio and visual streams. The figure shows lip movements congruent with the F0(2) vowels. The maximum opening of the mouth is centered on the cued audio vowel, but begins slightly before and remains somewhat after each cued auditory vowel. Each picture of the lips was extracted from the lip gesture movie that started 67 ms (i.e., two frames) before the corresponding auditory vowel and ended 100 ms after the offset of the auditory vowel (i.e., three frames).

approximately the time required for the build-up of auditory streaming (Bregman, 1978), participants were instructed to report the order in which vowels appeared by clicking on a graphical user interface displayed on the computer screen a few centimeters below the video of the lips. After participants confirmed their response, or at the end of the 10-s period during which the sequence was presented, there was silence for a period of 7 s, followed by the next sequence.

In one block, the 30 combinations of the three conditions of audiovisual congruence and the ten F0s were randomly repeated five times. Each participant ran eight blocks of trials. Overall, each combination was repeated 40 times (i.e., eight blocks × five repetitions). The experiment was divided into three sessions of 2 h each.

D. Results

Figure 3 shows percent-correct as a function of fundamental frequency and audio-visual (AV) congruence, averaged across participants. Responses were considered correct when the six vowels were reported in the correct order. A two-way repeated measures ANOVA was performed with fundamental frequency and AV congruence as repeated factors. Performance significantly decreased as the fundamental frequency increased [$F(9,81) = 22.13$; $p < 0.0001$], and AV congruence (visual condition) tended to reduce performance [$F(2,18) = 3.47$; $p = 0.05$].[1] There was no interaction between these factors [$F(18,162) = 0.87$; $p = 0.61$].[2] To clarify the effect of the AV congruence, Bonferroni-corrected post hoc tests were conducted. These tests revealed that the performance in C_0 was significantly better than in C_2 [$p = 0.05$]. There was no difference between C_1 and C_0 [$p = 0.37$] or between C_1 and C_2 [$p > 0.99$]. Note that the displayed standard error was across subject for each condition, while in the repeated measure ANOVA and in the post hoc tests, the error term was based on the difference between condition within subjects. Finally, it is worth noting that the same ANOVA performed when accepting responses with four vowels in the correct order (instead of six) strengthened the audiovisual effect (Effect of CV: $F(2,18) = 10.75$, $p < 0.001$; Effect of F0: $F(9,81) = 30$, $p < 0.0001$; Interaction: $F(18,162) = 1.35$, $p = 0.16$).

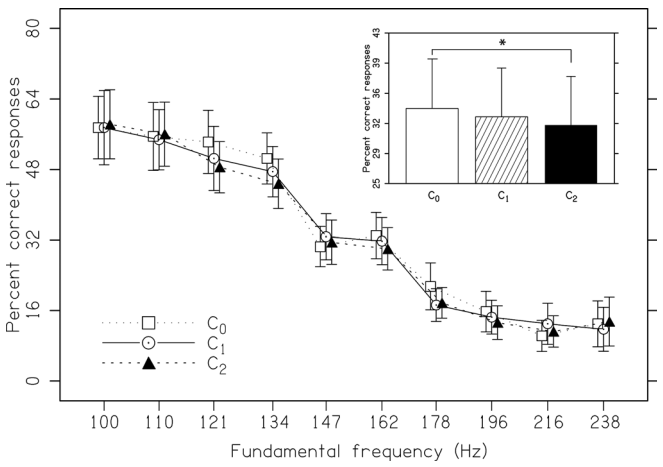

FIG. 3. Results from Experiment 1. Percent correct naming on the order of presentation of the six vowels is shown as a function of the F0 difference between alternate vowels and the visual condition. Bars represent the standard errors.

E. Discussion

The data from this experiment suggest that a phonetically congruent lip movement can enhance obligatory auditory streaming. In contrast, streaming was not significantly enhanced by lip-reading when the lip gesture consisted of a simple open−close gesture. One possibility is that the simple open/close rhythmic cue was not sufficient to elicit audiovisual interaction, which was weakened by the lack of phonetic congruence between the audio and visual inputs. This finding is consistent with the results reported by Rahne *et al.* (2007) and Rahne and Böckmann-Barthel (2009). These authors used basic geometric shapes of various sizes in synchrony with tones, which is similar to our condition C_1 in providing a visual rhythmic cue. Both the current results and those obtained by Rahne *et al.* (2008) suggest that a relatively high level of congruence between audio and visual signals may be required to observe an influence of audiovisual input on streaming.

The current data also showed that there was no statistical interaction between the effect of the visual cue and the effect of the fundamental frequency on streaming. This

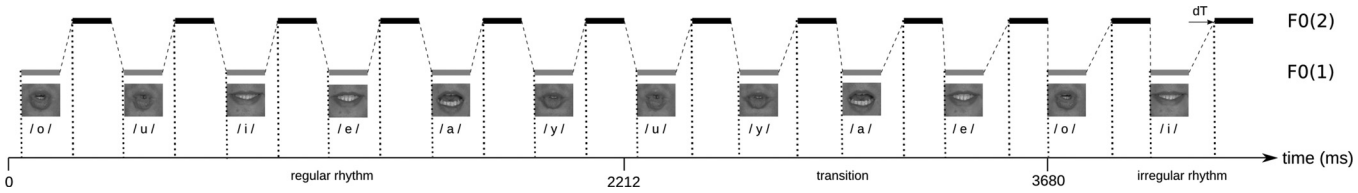

FIG. 4. Schematic representation of an audiovisual sequence. The sequence is first regular and then the audio-only stream (in black) is gradually delayed to reach the final value of dT that the listener has to detect. The audiovisual stream (in gray) is always regular. Vertical dotted lines exhibit the transition from a regular to an irregular rhythm of the audio-only stream.

finding suggests that these two factors contribute independently to primary streaming.

However, these conclusions need to be taken with caution because the task did not demonstrate as much sensitivity as expected and the reported effects are small. The small size of the effect could result in part from the fact that the audio and visual portions of the signal were generated independently. The degree of audiovisual congruence varied then from null (C_0) to only moderate (C_2) but remained much lower than in natural lip-reading situations. In Experiment 2, recorded vowels were employed in an attempt to strengthen congruence and confirm the role of audiovisual input on primary streaming.

III. EXPERIMENT 2

The results from the first experiment support the basic hypothesis that primary segregation is enhanced by lip-reading. However, the degree of audiovisual congruence was relatively low and might have accounted for the small (but significant) effect. To enhance the audiovisual coherence in the second experiment, all audiovisual vowels were presented from audio/video recordings of the same male speaker. This enhanced the synchrony between auditory and visual inputs, and provided more natural auditory and visual stimuli.

The task in this experiment was to detect a change in the presentation rate of vowel sequences alternating in pitch similar to those employed in the first experiment. Previous studies have shown that it is difficult to judge the relative timing of sounds that are perceived in different streams (Cusack and Roberts, 2000; Roberts *et al.,* 2002; Stainsby *et al.,* 2004). As in Experiment 1, conditions differing in congruence were created. In one condition, there was no AV congruence (C_0). In the other condition (C_3), alternate vowels were presented with an AV congruence stronger than that used in Experiment 1. As in the first experiment, poor performance in this task indicates that primary stream segregation has occurred. We predicted that the effect of lip-reading on segregation would be revealed by a lower performance in C_3 than in C_0.

A. Participants

Ten participants aged 18 to 24 (mean = 21.5, SD = 2.9) took part in Experiment 2. None had participated in Experiment 1. The audiometric threshold criteria were the same as in the first experiment.

B. Stimuli

As in Experiment 1, sequences of French vowels alternating in F0 were created. The same six French vowels were recorded (44.1 kHz, 16 bits) using simultaneous audio and video recording. The fundamental frequencies and durations of the natural productions were adjusted using STRAIGHT (Kawahara *et al.,* 1999) to reach a low and a high fundamental frequency and a fixed duration equal to 166 ms, including a 10 ms raised-cosine onset and offset ramp. The low fundamental frequency F0(1) was equal to 100 Hz. The second fundamental frequency F0(2) consisted of two possible values: 100 or 224 Hz. The stimuli were presented in intervals of twelve pairs of alternating vowels. The vowels were randomly chosen for each interval. The sequences started with either a F0(1) vowel or with a F0(2) vowel.

Two different lip gestures were used. In the first condition (C_0) of null AV congruence, the lips remained closed during the entire sequence. In the second condition (C_3) of strong AV congruence, natural lip gestures pronounced the odd-numbered (so alternate) auditory vowels in the sequence. The lip gestures for the six vowel sounds are displayed in Appendix B (Fig. 7). The lip gestures started 51 ms (i.e., two frames at 39 frames per second) before the corresponding vowel sound and finished 154 ms after (six frames at 39 frames per second). Figure 4 shows a schematic representation of a complete C_3 sequence. All the sequences were created on-line using a Python program. The stimuli were played diotically using a Sigmatel sound card connected to a Sennheiser HD 250 Linear II headphone. Visual stimuli were displayed on a monitor with a visual angle of approximately 6°. Listeners were comfortably seated in a double-walled, attenuated sound booth. The output level of the vowels was calibrated to 70 dB SPL (Larson Davis AEC101 and 824; American National Standards Institute, 1995).

C. Procedure

The method used in the current experiment was similar to that used by Cusack and Roberts (2000) and Roberts *et al.* (2002). A two-interval forced-choice method with a three-down, one-up decision rule (Levitt, 1971) was used to measure the smallest detectable temporal shift of the even numbered vowels (audio only) relative to the odd numbered vowels (audiovisual vowels). In each trial, the participants were required to identify the interval containing the temporal

shift, which was randomly assigned to one of the two intervals. The control sequence was an isochronous sequence of vowels, each separated by a constant 40 ms inter-stimulus interval (ISI). In the interval containing the target sequence, the first six pairs of vowels had a constant ISI of 40 ms to allow sufficient time for streaming to build up (Bregman, 1978). In the subsequent four pairs, the ISI between the vowels was progressively increased by an additional delay (rhythm deviation, dT) that ranged from 0 ms to a maximum value of 40 ms to avoid temporal overlap between two successive vowels. The dT reached after this transition phase was maintained during the last two pairs (see Fig. 4). The total duration of each sequence was 5 s. The silence between the two intervals was 7 s. It is worth noting that, because the audio-only vowels were delayed, the rhythm of the audiovisual vowel presentation remained constant in both intervals.

In the adaptive procedure, the initial value of dT was set to 20 ms. dT was then adjusted on a logarithmic scale. Specifically, the value of dT after each incorrect response was multiplied by 1.414, or divided by 1.414 after three successive correct responses. Each measurement continued until six reversals were reached. For the last four reversals, the step size was reduced to 1.189. A measurement was considered as saturated when 10 successive incorrect responses were provided with a dT value equal to 40 ms. The saturated measurements were assigned a threshold of 40 ms. If more than 50% of the measurements were saturated, the participant's data were not included in the analyses. If the measurement was not saturated, the geometric mean of the dT values for the last four reversals was used as the threshold estimate. Four such threshold estimates were made for each AV congruence and F0 condition, and the geometric mean of these estimates was used as the final value in the analyses. Each block consisted of four measurements (2 AV congruence $\times$ 2 F0) completed in a random order. An initial two blocks were considered training. A final four blocks were used to compute the individual thresholds for each condition.

D. Results

Two participants whose measures were saturated 75% and 62% of the time were not included in the analyses. Figure 5 shows the geometric mean of the dT thresholds across the eight remaining participants, as a function of the fundamental frequency and the AV congruence condition. As the measure is based on a detection threshold, the smaller the threshold, the better the performance. A two-way repeated measures ANOVA was performed with fundamental frequency and AV congruence condition as factors. The fundamental frequency had a detrimental effect on performance, which was consistent with the findings of Experiment 1 [$F(1,7) = 21.68$; $p < 0.01$]. Increased AV congruence also had a significant detrimental effect on performance, as indicated by larger dT thresholds [$F(1,7) = 5.85$; $p < 0.05$]. This detrimental effect of AV cues was consistent across seven out of eight listeners. The interaction between the two factors was not significant [$F(1,7) = 0.16$; $p = 0.71$].

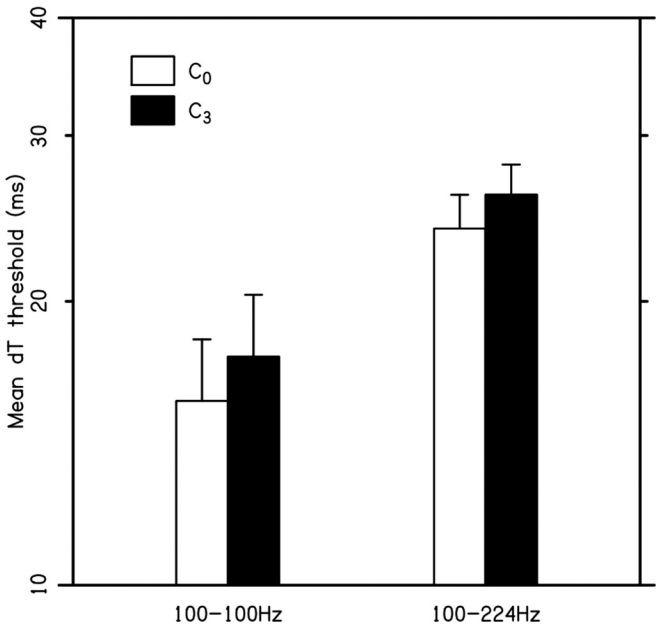

FIG. 5. Results of Experiment 2. The threshold for detecting a temporal offset between streams is plotted for each F0 difference and each visual condition across the eight participants. Errors bars represent one standard error.

E. Discussion

The results from Experiment 2 confirmed the hypothesis that more natural lip movements facilitated streaming relative to the no lip movement condition. This is consistent with the results of Experiment 1 in which a small effect of AV input on streaming was observed. The detection task used in the current experiment was likely more sensitive than that used in Experiment 1. In addition, the use of recorded audiovisual speech in the current study may have increased the audiovisual congruence and thereby strengthened the effect of the visual cue on auditory segregation. However, it should be noted that the AV stimuli in the current experiment were also altered. The natural AV desynchrony was reduced and the natural vowel durations were reduced to reach inter onset intervals that ensured that streaming would occur (van Noorden, 1975). As a consequence, assuming that naturalness or audiovisual congruence is required for maximal audiovisual integration, the modest size of the effect in the current experiment could be related to a congruence that was still lower than in natural situations. Finally, as in Experiment 1, the interaction between the visual cues and the F0 was not significant: the effect of adding a visual cue when the F0 condition induced streaming was not significantly different from that when the F0 condition did not induce streaming. This suggests that fundamental frequency difference and AV congruence act independently and may both contribute to the automatic segregation of competing speech.

IV. GENERAL DISCUSSION

A. Effect of lip gestures on obligatory streaming

The current findings suggest that part of the benefit of lip-reading in concurrent speech perception is a result of an

interaction between primary segregation and visual input. In fact, visual lip gestures that are phonetically congruent with the auditory speech signal are capable of inducing obligatory streaming. As suggested by Arnal *et al.* (2009) and Miller and D'Esposito (2005), who reported that two levels of interaction occur during the perception of AV speech, the interaction between auditory and visual input found in the current data could result from both low-level and high-level interactions.

As described in the Introduction, it is well established that streaming requires an initial build-up period during which the formation of different auditory streams occurs. In Experiment 1, subjects were locked out from responding during an initial 2-s period, to ensure that the percept had time to establish and stabilize. An interesting question not specifically addressed here involves the extent to which visual cues affect the time-course of streaming. The small effects observed here make a specific analysis of buildup difficult. However, the rhythm deviation task employed in Experiment 2 has been shown to be sensitive to buildup. Micheyl and Oxenham (2010) found that dT performance was related to the length of the sequence. Thus, it is possible to speculate that a portion of the overall differences observed as a result of visual condition in Experiment 2 could be attributed to differences in the rate at which streaming developed.

B. Audiovisual congruence

The small effect of an audiovisual signal on segregation suggested by Rahne *et al.* (2007) and Rahne and Böckmann-Barthel (2009) was observed in both Experiments 1 and 2. Previous studies showed that a one dB improvement in speech to noise ratio could correspond to a 5%–10% increase in intelligibility (Miller and Heise, 1950; Grant and Braida, 1991). Moreover, Grant and Seitz (2000) argued that a 2 dB AV effect on detection could be interpreted as a release from masking and could then lead to large increases in intelligibility. The small AV effect on streaming reported in Experiments 1 and 2 could then also potentially lead to a substantial effect on intelligibility in more realistic situations.

The tasks employed in Experiments 1 and 2 differed, making direct comparison difficult. However, the effect was larger in Experiment 2, where AV congruence was greater. The working hypothesis that audiovisual congruence of the stimuli is important for eliciting the effect of visual cues on primary auditory segregation is seemingly deserving of further attention. The congruence of the stimuli may depend on two factors. First, the temporal congruence between the amplitude envelope of the auditory input and the visual input appears to be important. One difference between our two experiments was the level of synchrony between the amplitude envelope of the auditory vowel input and the visual lip gestures. In Experiment 1, the auditory envelope remained at a constant level during the presentation of the lip gestures. In Experiment 2, the audio and visual parts of the signal were recorded simultaneously, and as such the audio temporal envelope was consistent with the lip gestures.

Second, the phonetic coherence between the auditory and visual inputs may have played a role in the congruence of the signals. In Rahne *et al.* (2007, 2008) and Rahne and Böckmann-Barthel (2009), the auditory signals were sequences of pure tones and the visual signals were geometric shapes. Thus, the signals did not have phonetic content or phonetic congruence. This may have been one reason why Rahne *et al.* observed no visual effect when streaming resulted from a clear acoustic cue. In Experiment 1, vowels were generated with the Klatt algorithm and the visual lip gestures were synthesized from a limited number of video frames. In Experiment 2, the vowels and the lip gestures were simultaneously recorded, and the audio stimuli were modified using STRAIGHT. Consequently, both the phonetic content and the phonetic coherence were greater in Experiment 2 than in Experiment 1. Although it is worth noting that the effect of visual cues might simply be task-dependent and was more easily elicited in Experiment 2 than in Experiment 1, the possible role of phonetic congruence in the effect of visual cues is in accord with Devergie *et al.* (2009), in which greater AV binding was found between auditory vowels and geometric shapes when the shapes changed in accordance with the natural lip gestures.

C. Neurophysiological correlates

The effect of visual lip gestures on primary streaming observed here is supported by neurophysiological research. Studies have reported that primary auditory cortical areas are sensitive to visual stimuli. For example, changes in a speech signal in the visual modality can be processed in the primary auditory cortex (Moettoenen *et al.*, 2002). This area of the brain can also be activated by a purely visual speech signal (Calvert *et al.*, 1997; Pekkola *et al.*, 2005). In summary, the current data provide support for a multisensory contribution (i.e., lip-reading) to low-level auditory processing (i.e., primary streaming) as reported in a recent review of neurophysiological work (Schroeder *et al.*, 2008).

D. Concluding remarks and perspectives

The results from these experiments confirmed the hypothesis that congruent lip gestures enhance primary streaming. The effect is consistent across listeners and experiments but remains small. Moreover, the degree of audiovisual coherence seems potentially related to the magnitude of this effect. In order to strengthen the effect of visual cues on primary streaming, futures studies could introduce more naturalness (natural synchrony, natural duration) and vowel-consonant-vowel (VCV) sequences with more lip kinematics to enhance coherence, focus on F0 values leading to bistable percepts and compare the duration of build-up in various audiovisual coherence conditions either psychophysically or with EEG recording.

ACKNOWLEDGMENTS

This work was supported by grants from the Région Rhones-Alpes Auvergne Cluster HVN 2007, the Agence Nationale de Recherche (ANR-08-BLAN-0167-01), and the National Institute on Deafness and Other Communication Disorders (R01 DC08594). We thank the participants of this study, as well as Christophe Savariaux for his help in recording the audiovisual material.

APPENDIX A

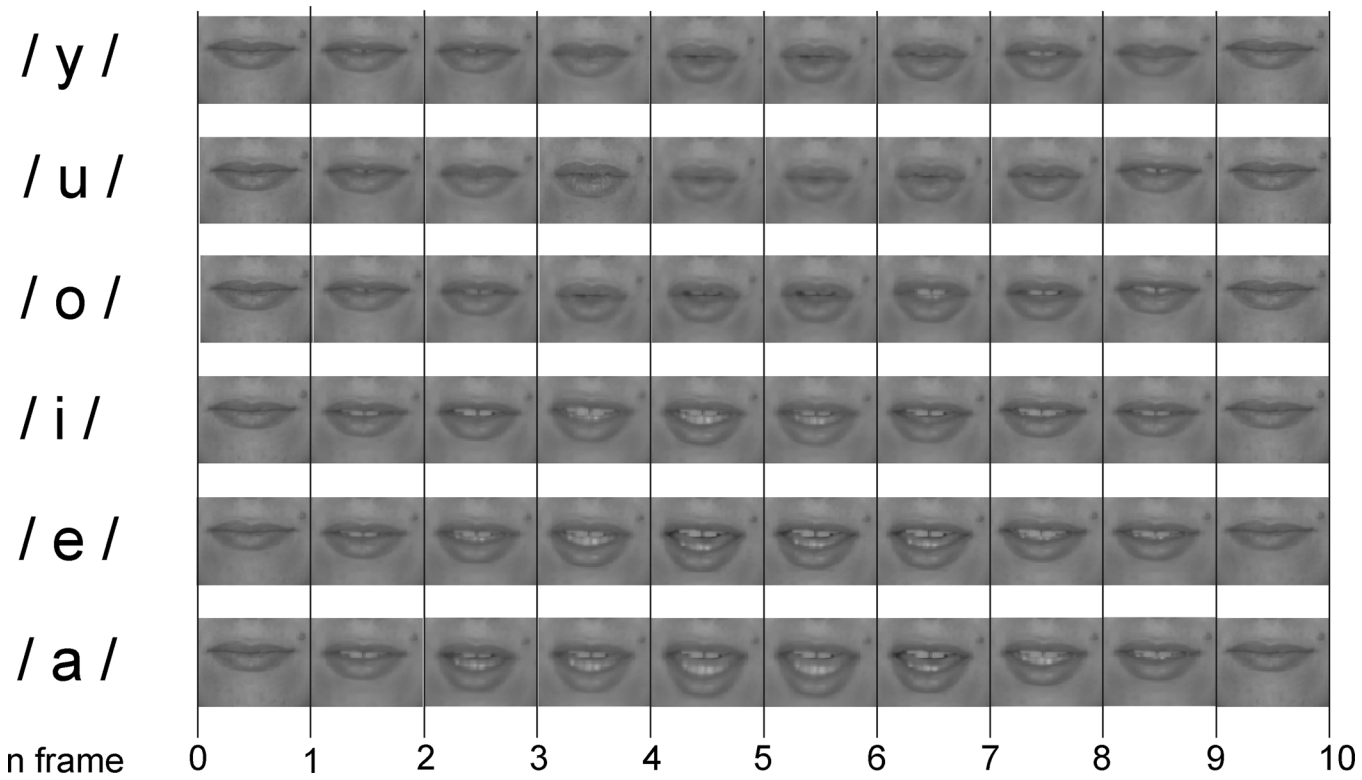

FIG. 6. Visual lip gestures in Experiment 1.

APPENDIX B

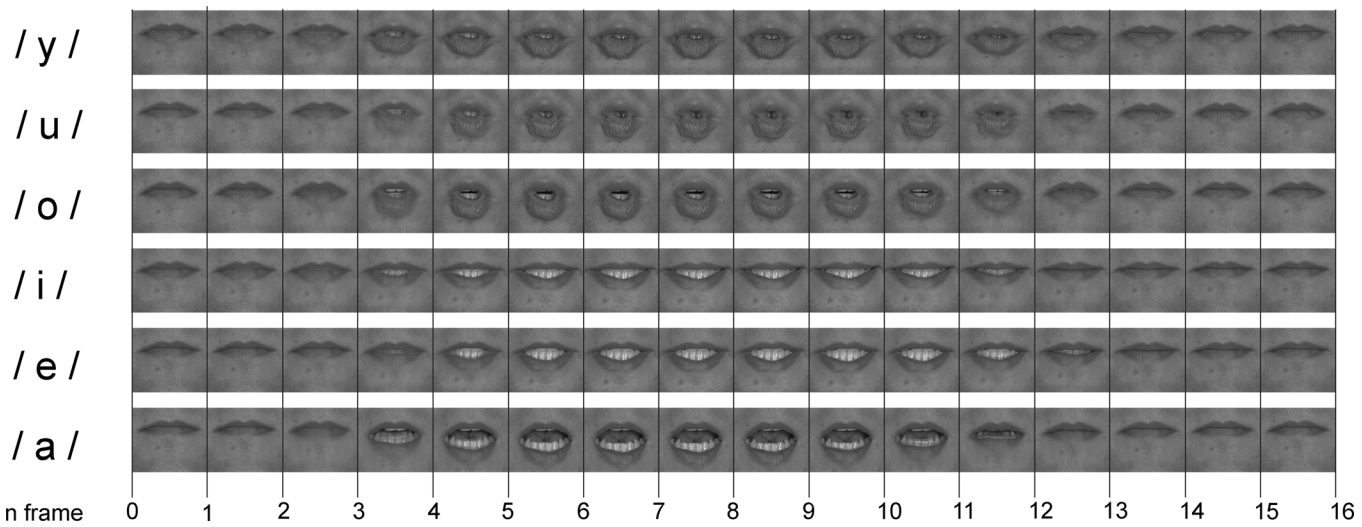

FIG. 7. Visual lip gestures in Experiment 2.

[1]An ANOVA focusing on only the central F0 range (between 121 and 196 Hz) which can be assumed to elicit bistable auditory organization, revealed a larger effect of AV congruence [$F(2,18) = 8.55$, $p < 0.005$].

[2]A Geisser Greenhouse correction was also performed to compensate for the lack of sphericity and provided the same pattern of results [F0: ε;(1.89, 17.02) = 0.21, $p < 0.0001$, AV congruence: ε;(1.32, 11.86) = 0.65, $p = 0.08$, interaction ε;(5.26, 47.39) = 0.29, $p = 0.51$].

American National Standards Institute (**1995**). *ANSI S3.7-R2003: Methods for Coupler Calibration of Earphones*, American National Standards Institute, NY.

American National Standards Institute (**2004**). *ANSI S3.21-2004: Methods for Manual Pure-Tone Threshold Audiometry*, American National Standards Institute, NY.

Arnal, L. H., Morillon, B., Kell, C. A., and Giraud, A.-L. (**2009**). "Dual neural routing of visual facilitation in speech processing." J. Neurosci. **29**, 13445–13453.

Bernstein, L. E., Auer, E. T. J., and Takayanagi, S. (**2004**). "Auditory speech detection in noise enhanced by lipreading," Speech Commun. **44**, 5–18.
Berthommier, F. (**2003**). "A phonetically neutral model of the low-level audiovisual interaction," in *Proceedings of the International Conference on Audio-Visual Speech Processing*, 89–94 (Institut de la Communication Parlée, St. Jorioz, France).
Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., and Giard, M.-H. (**2008**). "Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans." J. Neurosci. **28**, 14301–14310.
Brancazio, L. and Brancazio, L. (**2004**). "Lexical influences in audiovisual speech perception." J. Exp. Psychol. Hum. Percept. Perform. **30**, 445–463.
Bregman, A. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sounds* (MIT Press, Cambridge, MA), 1–773.
Bregman, A. S. (**1978**). "Auditory streaming is cumulative." J. Exp. Psychol. Hum. Percept. Perform. **4**, 380–387.
Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W., Iversen, S. D., and David, A. S. (**1997**). "Activation of auditory cortex during silent lipreading," Science **276**, 593–596.
Cusack, R., and Roberts, B. (**2000**). "Effects of differences in timbre on sequential grouping," Percept. Psychophys. **62**, 1112–1120.
Devergie, A., Berthommier, F., and Grimault, N. (**2009**). "Pairing audio speech and various visual displays: binding or not binding?" in *Proceedings of the International Conference on Audio-Visual Speech Processing* (School of Computing Sciences, Norwich, UK), pp. 140–144.
Gaudrain, E., Grimault, N., Healy, E. W., and Béra, J.-C. (**2007**). "Effect of spectral smearing on the perceptual segregation of vowel sequences," Hear. Res. **231**, 32–41.
Gaudrain, E., Grimault, N., Healy, E. W., and Béra, J.-C. (**2008**). "Streaming of vowel sequences based on fundamental frequency in a cochlear-implant simulation," J. Acoust. Soc. Am. **124**, 3076–3087.
Grant, K. W., and Braida, L. D. (**1991**). "Evaluating the articulation index for auditory- visual input," J. Acoust. Soc. Am. **89**, 2952–2960.
Grant, K. W., and Walden, B. E. (**1996**). "Spectral distribution of prosodic information," J. Speech Hear. Res. **39**, 228–238.
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**, 1197–1208.
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., van Wassenshove, V., and Poeppel, D. (**2003**). "Discrimination of auditory-visual synchrony," in *Proceedings of the International Conference on Audio-Visual Speech Processing* (Institut de la Communication Parlée, St jorioz, France), pp. 31–35.
Grant, K. W., Wassenhove, V., and Poeppel, D. (**2004**). "Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony," Speech Commun. **44**, 43–53.
Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time-frequency-smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**, 187–207.
Kayser, C., Petkov, C. I., and Logothetis, N. K. (**2008**). "Visual modulation of neurons in auditory cortex," Cereb. Cortex **18**, 1560–1574.
Klatt, D. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.
Massaro, D. W., and Cohen, M. M. (**1983**). "Evaluation and integration of visual and auditory information in speech perception," J. Exp. Psychol. Hum. Percept. Perform. **9**, 753–771.
Micheyl, C. and Oxenham, A. J. (**2010**). "Objective and subjective psychophysical measures of auditory stream integration and segregation," J. Assoc. Res. Otolaryngol., http://dx.doi.org/10.1007/s10162-010-0227-2 (date last viewed 10/27/10).
Micheyl, C., Tian, B., Carlyon, R. P., and Rauschecker, J. P. (**2005**). "Perceptual organization of tone sequences in the auditory cortex of awake macaques," Neuron **48**, 139–148.
Miller, G. A., and Heise, G. A. (**1950**). "The thrill threshold," J. Acoust. Soc. Am. **22**, 637–638.
Miller, L. M. and D'Esposito, M. (**2005**). "Perceptual fusion and stimulus coincidence in the cross-modal integration of speech," J. Neurosci. **25**, 5884–5893.
Moettoenen, R., Krause, C. M., Tiipana, K., and Sams, M. (**2002**). "Processing of changes in visual speech in the human auditory cortex," Cognit. Brain Res. **13**, 417–425.
Moore, B. C. J., and Gockel, H. (**2002**). "Factors influencing sequential stream segregation," Acta Acust. **88**, 320–333.
Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., and Sams, M. (**2005**). "Primary auditory cortex activation by visual speech: an fmri study at 3 t," Neuroreport **16**, 125–128.
Pressnitzer, D., Sayles, M., Micheyl, C., and Winter, I. M. (**2008**). "Perceptual organization of sound begins in the auditory periphery," Curr. Biol. **18**, 1124–1128.
Rahne, T., Böckmann, M., von Specht, H., and Sussman, E. S. (**2007**). "Visual cues can modulate integration and segregation of objects in auditory scene analysis," Brain. Res. **1144**, 127–135.
Rahne, T., and Böckmann-Barthel, M. (**2009**). "Visual cues release the temporal coherence of auditory objects in auditory scene analysis," Brain. Res. **1300**, 125–134.
Rahne, T., Deike, S., Selezneva, E., Brosch, M. König, R., Scheich, H., Böckmann, M., and Brechmann, A. (**2008**). "A multilevel and cross-modal approach towards neuronal mechanisms of auditory streaming," Brain. Res. **1220**, 118–131.
Roberts, B., Glasberg, B. R., and Moore, B. C. J. (**2002**). "Primitive stream segregation of tone sequences without differences in fundamental frequency or passband," J. Acoust. Soc. Am. **112**, 2074–2085.
Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (**2008**). "Neuronal oscillations and visual amplification of speech," Trends Cogn. Sci. **12**, 106–113.
Stainsby, T. H., Moore, B. C., and Glasberg, B. R. (**2004**). "Auditory streaming based on temporal structure in hearing-impaired listeners," Hear. Res. **192**, 119–130.
Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.
van Noorden, L. (**1975**). "Temporal coherence in the perception of tone sequences," Ph.D. dissertation, Technische Hogeschool Eindhovern, Eindhoven, The Netherlands.
van Wassenhove, V., Grant, K., and Poeppel, D. (**2005**). "Visual speech speeds up the neural processing of auditory speech," PNAS **102**, 1181–1186.