Available online at www.sciencedirect.com
SCIENCE DIRECT
Neuroscience Letters 393 (2006) 40–44
Neuroscience Letters
www.elsevier.com/locate/neulet

Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task

Argiro Vatakis *, Charles Spence

Crossmodal Research Laboratory, Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, UK

Received 16 May 2005; received in revised form 31 August 2005; accepted 14 September 2005

Abstract

This study investigated people's sensitivity to audiovisual asynchrony in briefly-presented speech and musical videos. A series of speech (letters and syllables) and guitar and piano music (single and double notes) video clips were presented randomly at a range of stimulus onset asynchronies (SOAs) using the method of constant stimuli. Participants made unspeeded temporal order judgments (TOJs) regarding which stream (auditory or visual) appeared to have been presented first. The accuracy of participants' TOJ performance (measured in terms of the just noticeable difference; JND) was significantly better for the speech than for either the guitar or piano music video clips, suggesting that people are more sensitive to asynchrony for speech than for music stimuli. The visual stream had to lead the auditory stream for the point of subjective simultaneity (PSS) to be achieved in the piano music clips while auditory leads were typically required for the guitar music clips. The PSS values obtained for the speech stimuli varied substantially as a function of the particular speech sound presented. These results provide the first empirical evidence regarding people's sensitivity to audiovisual asynchrony for musical stimuli. Our results also demonstrate that people's sensitivity to asynchrony in speech stimuli is better than has been suggested on the basis of previous research using continuous speech streams as stimuli.

Keywords: Synchrony perception; TOJ; Speech; Music; Audition; Vision

Watching a live satellite news report on the television provides one of the many everyday examples where a temporal mismatch can be detected between what one sees and hears, in this case between the sight of the reporters' lips moving and the sound of his/her voice. The broadcasting industry, aware of the fact that people are sensitive to asynchrony in speech stimuli, has established a maximum acceptable asynchrony in broadcasting, stating that the auditory signal should not lead by more than 45 ms or else lag by more than 125 ms [10]. Research suggests that within this temporal window only a minimal deterioration in program intelligibility will be observed (cf. [19]). However, one important, but as yet unanswered, question regards how sensitive people are to the asynchrony of speech versus to other kinds of complex non-speech stimuli, such as, for example, musical instruments? Despite the recent increase of interest in the multisensory perception of synchrony (e.g., [13,22]), the majority of research in this area has tended to focus on the perception of synchrony and asynchrony for simple transitory stimuli (such as brief noise bursts, light flashes, and punctate tactile stimuli; e.g., [8,21,28,29]).

One of the first studies to investigate the perception of synchrony for speech stimuli was reported by Dixon and Spitz [5]. Participants in their study had to monitor videos that started in synchrony and were gradually desynchronized at a rate of 51 ms/s (up to a maximum asynchrony of 500 ms) with either the auditory or visual stream leading. The participants had to respond as soon as they detected that the videos were asynchronous. Dixon and Spitz found that the auditory stream had to lag by an average of 258 ms or lead by 131 ms before the asynchrony was detected. More recently, Grant et al. [7] reported that participants only noticed the asynchrony in a continuous stream of audiovisual speech when the speech sounds led the visual lip movements by at least 50 ms or else lagged by 220 ms or more (see also [9]).

In order to understand speech processing and its potentially 'special' nature (e.g., [2,14,26,32] Munhall and Vatikiotis-Bateson, 2004) it may be informative to compare the perception of synchrony in speech under conditions of continuous (i.e., uttering sentences) versus abrupt (using syllables) speech perception. It will also be of interest to study the perception of

* Corresponding author. Tel.: +44 1865 271307; fax: +44 1865 310447.
E-mail address: argiro.vatakis@psy.ox.ac.uk (A. Vatakis).

0304-3940/$ – see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.neulet.2005.09.032

synchrony for other complex non-speech events. However, only very limited evidence currently exists regarding TOJs in speech (continuous or abrupt) and other complex non-speech events, such as music or object-actions. In Dixon and Spitz's [5] study, participants were not only presented with speech videos but also with an object-action video (the action of a hammer hitting a peg). Audiovisual asynchrony was detected more rapidly when viewing the object-action event than when viewing the speech stimuli. Specifically, an auditory lag of 188 ms lag or a lead of 75 ms was required for the detection of asynchrony in the object-action videos. Participants therefore appeared to be more sensitive to the presence of asynchrony in the object-action videos than in the speech videos (see also [9]).

What are the factors that account for the differences in temporal processing acuity reported in studies with simple, transitory stimuli and those using more complex speech and non-speech events? The audiovisual asynchrony values reported in previous studies (e.g., [5,7] see also [9]) might not actually provide an accurate reflection of people's sensitivity to asynchrony in audiovisually-presented complex stimuli due to a number of methodological factors. For example, in both Dixon and Spitz and Grant et al.'s studies, the auditory stimuli were presented over headphones while the visual stimuli were presented from in front of the participant on a monitor. Recent studies have shown that the integration of multisensory stimuli can be facilitated by their spatial coincidence (e.g., see [3]; though see also [16]); hence, previous studies may have systematically overestimated people's sensitivity to asynchrony by providing additional spatial cues to temporal asynchrony that are not present when auditory and visual stimuli come from the same location, as when we listen to someone speaking in real life. Additionally, it is unclear whether the gradual desynchronization of the videotapes in Dixon and Spitz's study might inadvertently have presented participants with subtle auditory pitch-shifting cues (see [18]). Finally, the fact that no catch trials were presented in Dixon and Spitz's study means that the influence of criterion shifting on performance cannot be assessed. These factors, should they prove important for temporal perception, would predict that TOJ performance for speech stimuli may actually be worse than has been reported in the literature to date.

Previous research in this area has tended to focus on speech events while ignoring other equally complex events, such as music. However, music might serve as a better stimulus (than object-actions and simple sound bursts and light flashes) for comparison with speech, given its complex time-varying nature, and the fact that listening to both speech and music has been shown to activate a number of the same neural structures (e.g., [31] though see [30]). It is therefore somewhat surprising to find that people's sensitivity to the temporal synchrony of musical videos has never been studied previously (though see [24], for an anecdotal report on this issue). Instead, speech has typically been adopted as the most crucial (if not the only), ecologically valid stimulus in studies of audiovisual perception for complex stimuli (e.g., [23]).

In the present study, we examined people's sensitivity to audiovisual asynchrony using both complex speech and musical stimuli. We focused on brief video clips of speech stimuli consisting of phonemes and syllables, and brief musical stimuli consisting of single and double notes. The perception of synchrony was assessed for the speech and musical stimuli using a temporal order judgment (TOJ) task with varied stimulus onset asynchronies (SOAs) using the method of constant stimuli. It should be noted that while a number of previous studies have examined the consequences of introducing audiovisual asynchrony on speech perceptibility (e.g., [14,18,19]), no one has used a TOJ task to assess sensitivity to asynchrony for either speech or musical stimuli before.

Twenty-one participants (10 male and 11 female) aged between 19 and 33 years (mean age of 24 years) were given a £5 (UK Sterling) gift voucher or course credit in return for taking part in the experiment. All of the participants were naïve as to the purpose of the study, none had any prior musical training, and all reported having normal or corrected-to-normal hearing and visual acuity. The experiment took approximately 50 min to complete.

The experiment was conducted in a completely dark sound-attenuated booth. During the experiment, the participants were seated comfortably at a small table facing straight-ahead. The visual stimuli were presented on a 17-in. (43.18 cm) TFT colour LCD monitor (SXGA 1240×1024 pixels resolution; 60-Hz refresh rate), placed at eye level, 68 cm in front of the participants. The auditory stimuli were presented by means of two Packard Bell Flat Panel 050 PC loudspeakers, one placed 25.4 cm to either side of the centre of the monitor. The audiovisual stimuli consisted of twelve video clips presented on a black background, using the Presentation programming software (Neurobehavioral Systems, Inc., CA). The video clips (400×300-pixel, Cinepak Codec video compression, 16-bit Audio Sample Size, 24-bit Video Sample Size, 25 frames/s) were processed using the Adobe Premiere 6.0 software package. The video clips consisted of the following: (a) the face of a British male, looking directly at the camera, and saying /a/ and /p/ (both videos were 967 ms long); (b) the same male uttering the syllables /lo/ and /me/ (833 ms duration); (c) a male playing the musical notes "a" and "d" on a classical guitar (only the centre of the body of the guitar was visible; 1700 ms duration); (d) the same male playing the combinations of notes "db" and "eg" on a classical guitar (2200 ms duration); (e) a bird's-eye view of the hands of a female playing the notes "a" and "d" on the piano (1700 ms duration); (f) the same female playing the note combinations "ce" and "fd" on the piano (2200 ms duration).

At the beginning and end of each video clip, a still image (extracted from the first and last 33.33 ms of each clip) and background acoustic noise was presented for a duration equivalent to the SOA (the SOA values are reported below) in order to avoid cuing the participants as to the nature of the audiovisual delay. In order to achieve a smooth transition at the start and end of each video clip, a cross-fading effect of 33.33 ms was added between the still image and the video clip. Therefore, each video started with participants viewing a black screen, the centre of the screen gradually transitioned to the still image and subsequently transitioned to the video clip (a similar transition was added to the end of each video clip). The participants responded using a standard computer keyboard, pressing the "V" key to indicate that the

visual stream appeared to have been presented first, and the "A" key to indicate that the auditory stream appeared to have been presented first.

Nine SOAs between the auditory and visual stimuli were used: ±400, ±300, ±200, ±100, and 0 ms. Negative SOAs indicate that the auditory stream was presented first, whereas positive values indicate that the visual stream was presented first. The participants completed one block of 12 practice trials before the main experimental session in order to familiarize themselves with the task and the video clips. The practice trials were followed by 10 blocks of 108 experimental trials, consisting of one presentation of each of the 12 video clips at each of the 9 SOAs in each block of trials. The various SOAs were presented in a random order with the sole restriction that a given video clip was not presented on consecutive trials.

Before the start of the experiment, the experimenter gave a detailed verbal description of the task to participants and they were allowed to ask any clarificatory questions. The participants were asked about their prior experience of music and any previous training they might have had with musical instruments. At the start of the experiment, the participants were informed that they would have to decide on each trial whether the auditory or visual video stream appeared to have been presented first, and that they would sometimes find this difficult, in which case they should make an informed guess as to the order of stimulus presentation. The participants were also informed that the task was self-paced, and that they should respond only when confident of their response. The participants were told that they did not have to wait until the video clip had finished before making their response, but that a response had to be made before the experiment would advance to the next trial. At the beginning of each block of trials the word "READY" was presented on the screen and the participants had to press the "ENTER" key to start the block. The participants were instructed prior to the experiment not to move their heads and to maintain their fixation on the centre of the monitor throughout each block of trials. The participants were allowed to take breaks between the blocks of experimental trials.

The proportions of 'vision first' responses were converted to their equivalent *z*-scores under the assumption of a cumulative normal distribution [6]. Data from the nine SOAs were used to calculate best-fitting straight lines for each participant for each condition, which, in turn, were used to derive values for the slope and intercept. These two values were used to calculate the just noticeable difference (JND = 0.675/slope; since ±0.675 represents the 75 and 25% point on the cumulative normal distribution) and the point of subjective simultaneity (PSS = −intercept/slope) values (see [4], for further details). The JND provides a standardized measure of the accuracy with which participants could judge the temporal order of the auditory and visual stimuli. The PSS indicates the amount of time by which one sensory modality had to lead the other in order for synchrony to be perceived (i.e., for participants to make the 'sound first' and 'vision first' responses equally often). For all of the analyses reported here, Bonferroni-corrected *t*-tests (where $p < .05$ prior to correction) were used for all post-hoc comparisons. The JND and PSS data for each of the speech, piano and guitar categories were initially analysed using one-way analysis of variance (ANOVA) with the factor of Category Exemplar (four levels; i.e., for speech: /a/, /p/, /lo/, and /me/; see Table 1).

Analysis of the JND data revealed no significant main effect of Category Exemplar [$F < 1$ for speech, guitar, and piano], showing that there was no difference in the accuracy of participant's TOJ responses as a function of the Category Exemplar within each of the three Stimulus Types. (Note also that a preliminary analysis of the data revealed no differences in performance for the short versus long videos; all $F < 1$.) In order to compare TOJ performance as a function of Stimulus Type, the data were averaged over the Category Exemplar factor. A one-way ANOVA performed with the factor of Stimulus Type (speech, guitar music, and piano music), revealed a significant main effect [$F(2,251) = 8.33$, $p < .01$] (see Fig. 1A), showing that temporal discrimination performance was significantly better for speech stimuli than for either the guitar or piano music stimuli [both *t*-test comparisons $p \leq .01$] (see Table 1). Subsequent analysis of the data with block number as an additional factor revealed no effect of practice on performance [$F < 1$] (though see [24]).

Analysis of the PSS data revealed a significant main effect of Category Exemplar [$F(3,83) = 7.69$, $p < .01$] for the speech stimuli (see Fig. 1B), but not for either the guitar or piano music [both $F < 2$]. The mean PSS values for the guitar and piano music were −7 ms and +41 ms, respectively. The speech stimulus /a/ requiring an auditory lead of 66 ms for the PSS to be achieved which was significantly different from the auditory lead of 8 ms required for /p/ stimulus, or the visual leads of 19 ms and 27 ms required for the syllables /lo/ and /me/, respectively [$p = .05$, $p < .01$, and $p < .01$, for the *t*-test comparisons]. The only PSS values to differ significantly from objective simultaneity (i.e., for 0 ms) were obtained for the speech letter /a/, and for the piano notes 'd' and 'ce' [$t(20) = -4.56$, $p < .01$; $t(20) = 2.58$, $p = .02$; and $t(20) = 3.02$, $p = .01$, respectively].

The results of the present study provide the first empirical evidence regarding people's sensitivity to asynchrony in musical

Table 1
Mean JND and PSS values (in ms), and their standard errors, derived from the TOJ task for the speech and music video clips as a function of the speech sound or musical note(s) presented

Condition	Category Exemplar	JND		PSS	
		Mean	S.E.	Mean	S.E.
Speech	a	101	7.0	−66	14.5
	p	94	6.3	−8	14.1
	lo	95	5.2	19	16.0
	me	95	5.0	27	16.3
Guitar	a	109	7.2	12	21.4
	d	104	10.3	−34	18.1
	db	116	11.0	−7	15.1
	eg	133	13.0	−23	16.7
Piano	a	110	9.0	18	18.0
	d	124	10.4	37	14.5
	ce	116	9.1	47	15.5
	fd	135	13.0	61	30.0

Negative PSS values indicate that the auditory stream had to lead for synchrony to be perceived.

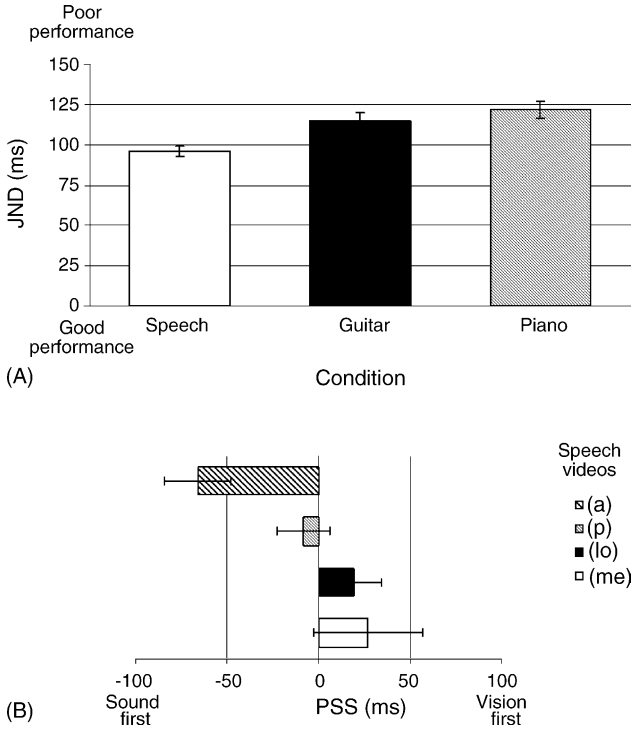

Fig. 1. (A) Averaged JNDs for the stimulus categories of speech, guitar music, and piano music. (B) PSSs for the four speech video clips. The error bars represent the standard errors of the mean.

stimuli. The results show that people were better able to detect the temporal asynchrony present in the desynchronized audiovisual speech videos than in either the guitar or piano music video clips (see Fig. 1A). Nevertheless, the JND values reported for speech stimuli in the present study were still noticeably higher than those observed in previous studies that have used simple sound-light pairs as experimental stimuli (e.g., [8,29] though see also [17]). However, perhaps more importantly, the JNDs reported for the brief speech stimuli were much smaller than those observed in previous studies of asynchrony detection using continuous speech stimuli [5,7]. The auditory speech stream had to lag the visual stream by more than 250 ms or else lead by more 130 ms before the asynchrony became noticeable in Dixon and Spitz's study, and to lag by 220 ms or lead by 50 ms in Grant et al.'s study. The lower values reported in the present study may therefore reflect the fact that we used only brief stimuli with the controlled viewing of just the area around the speakers' mouth, and minimal head movements [25]. Alternatively however, they may also reflect the fact that the TOJ task used here offers a more sensitive index of people's sensitivity to asynchrony than the tasks used in previous studies.

Analysis of the PSS data revealed a greater variability in modality lead/lag for the speech videos than for the musical videos. For example, while the auditory stream had to lead by 66 ms for the /a/ speech stimulus, the visual stream had to lead by 27 ms for the /me/ speech stimulus (see Fig. 1B). This difference may reflect the fact that the phonetic and physical properties involved in the production of speech sounds vary as a function of the speech sound being uttered [12]. So, for example, the production of the vowel /a/ depends on the position of the jaw, tongue, and lips and for the bilabial, nasal consonant /m/ closure of the oral cavity is required, while uttering /ma/ requires the rapid combination of those movements. These differences in the nature of the production of different speech sounds may explain the variations in the auditory or visual delays/leads required for the successful perception of synchrony for audiovisual speech [15]. They may also help to account for the decreased sensitivity to continuous speech reported in previous studies, where the PSS will presumably have varied continuously depending on the particular speech sound being uttered at any moment. By contrast, the plucking of a guitar chord or the striking of a piano key does not exhibit the same physical differences in the relative timing of audition and vision as a function of the particular note played (except perhaps in the case of anticipatory movements, where preparatory finger movements might provide the observer with misleading timing information; [1]).

It has been argued that speech and music share a number of important properties. For instance, they are both composed of perceptually discrete basic elements (such as phonemes and notes) which can be organized into meaningful, syntax-governed sequences (such as phrases and tunes). Interestingly, however, the results reported here reveal better temporal discrimination accuracy for speech stimuli than for musical stimuli. This might be due to the greater prior experience that people have with the perception of audiovisual speech stimuli (note that the participants in our study had no prior musical experience) thus perhaps making the speech stimuli more perceptually salient to our participant's than the musical stimuli. It might even be due to the putatively 'special' nature of speech processing (e.g., see [2,14,26,32]). However, an alternative account for these findings relates to possible differences in the temporal profile of the stimuli (i.e., the rise times for speech and musical stimuli may be different which might also affect TOJ performance; cf. [11]). Finally, the higher sensitivity observed for asynchrony in speech may also be related to recent reports emerging from mirror neuron studies on speech perception, where left hemisphere excitability of the motor units underlying speech production have been observed when participants either listen to speech without any visual stimulation, or even by the mere sight of visual speech-related lip movements with no auditory stimulation [27].

Recent studies have shown that several commonalities and differences in the activations of various brain structures when people process speech or musical stimuli [30,31]. To our knowledge, however, no previous research had been carried out using musical instruments as an experimental stimulus for studying people's sensitivity to asynchrony. Future combined psychophysical and neuroimaging studies will therefore help to further our understanding of the brain's processing of speech and music in more detail, and thus to promote a better understanding of the mechanisms involved in the multisensory perception of synchrony for complex realistic stimuli (cf. [17,20]).

References

[1] A.P. Baader, O. Kazennikov, M. Wiesendanger, Coordination of bowing and fingering in violin playing, Cogn. Brain Res. 23 (2005) 436–443.

[2] L.E. Bernstein, E.T. Auer, J.K. Moore, Audiovisual speech binding: convergence or association? in: G.A. Calvert, C. Spence, B.E. Stein (Eds.), The Handbook of Multisensory Processing, MIT Press, Cambridge, MA, 2004, pp. 203–223.
[3] G.A. Calvert, C. Spence, B.E. Stein (Eds.), The Handbook of Multisensory Processing, MIT Press, Cambridge, MA, 2004.
[4] S. Coren, L.M. Ward, J.T. Enns, Sensation & Perception, sixth ed., Harcourt Brace, Fort Worth, 2004.
[5] N.F. Dixon, L. Spitz, The detection of auditory visual desynchrony Perception 9 (1980) 719–721.
[6] D.J. Finney, Probit Analysis: Statistical Treatment of the Sigmoid Response Curve, Cambridge University Press, London, UK, 1964.
[7] K.W. Grant, V. van Wassenhove, D. Poeppel, Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony, J. Acoust. Soc. Am. 108 (2004) 1197–1208.
[8] I.J. Hirsh, C.E. Sherrick Jr., Perceived order in different sense modalities, J. Exp. Psychol. 62 (1961) 424–432.
[9] M.P. Hollier, A.N. Rimell, An experimental investigation into multimodal synchronisation sensitivity for perceptual model development, 105th AES Convention, 1998, Preprint No. 4790.
[10] ITU-R BT.1359-1, Relative timing of sound and vision for broadcasting (Question ITU-R 35/11), 1998.
[11] P. Jaśkowski, Temporal-order judgement and reaction time to stimuli of different rise times, Perception 22 (1993) 963–970.
[12] R.D. Kent, The Speech Sciences, Singular, San Diego, CA, 1997.
[13] A.J. King, Multisensory integration: strategies for synchronization, Curr. Biol. 15 (2005) R339–R341.
[14] D.W. Massaro, From multisensory integration to talking heads and language learning, in: G.A. Calvert, C. Spence, B.E. Stein (Eds.), The Handbook of Multisensory Processing, MIT Press, Cambridge, MA, 2004, pp. 153–176.
[15] V. van Wassenhove, K.W. Grant, D. Poeppel, Temporal window of integration in bimodal speech perception, J. Cogn. Neurosci., submitted for publication.
[16] M.M. Murray, S. Molholm, C.M. Michel, D.J. Heslenteld, W. Ritter, D.C. Javitt, C.E. Schroeder, J.J. Foxe, Grabbing your ear: rapid auditory-somatosensory multisensory interactions in low-level sensory cortices are not constrained by stimulus alignment, Cereb. Cortex 15 (2005) 963–974.
[17] J. Navarra, A. Vatakis, M. Zampini, W. Humphreys, S. Soto-Faraco, C. Spence, Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. Cogn. Brain Res., in press.
[18] B. Reeves, D. Voelker, Effects of audio-video asynchrony on viewer's memory, evaluation of content and detection ability. Research report prepared for Pixel Instruments. Los Gatos, California, 1993.
[19] S. Rihs, The influence of audio on perceived picture quality and subjective audio-visual delay tolerance, in: R. Hamberg, H. de Ridder (Eds.), Proceedings of the MOSAIC Workshop: Advanced Methods for the Evaluation of Television Picture Quality, vol. 133–137, Eindhoven, 18–19 September 1995.
[20] H.M. Saldaña, L.D. Rosenblum, Visual influences on auditory pluck and bow judgments, Percept. Psychophys. 54 (1993) 406–416.
[21] C. Spence, D.I. Shore, R.M. Klein, Multisensory prior entry, J. Exp. Psychol.: Gen. 130 (2001) 799–832.
[22] C. Spence, S.B. Squire, Multisensory integration: maintaining the perception of synchrony, Curr. Biol. 13 (2003) R519–R521.
[23] Q. Summerfield, Some preliminaries to a comprehensive account of audio-visual speech perception, in: B. Dodd, R. Campbell (Eds.), Hearing by Eye: The Psychology of Lip-reading, LEA, London, 1987, pp. 3–51.
[24] Q. Summerfield, Lipreading and audio-visual speech perception, Philos. Trans. R. Soc. Lond. B 335 (1992) 71–78.
[25] S.M. Thomas, T.R. Jordan, Contributions of oral and extraoral facial movement to visual and audiovisual speech perception, J. Exp. Psychol.: Hum. Percept. Perform. 30 (2004) 873–888.
[26] J. Tuomainen, T.S. Andersen, K. Tiippana, M. Sams, Audio-visual speech is special, Cognition 96 (2005) B13–B22.
[27] K.E. Watkins, A.P. Strafella, T. Paus, Seeing and hearing speech excites the motor system involved in speech production, Neuropsychologia 41 (2003) 989–994.
[28] M. Zampini, S. Guest, D.I. Shore, C. Spence, Audio-visual simultaneity judgments, Percept. Psychophys. 67 (2005) 531–544.
[29] M. Zampini, D.I. Shore, C. Spence, Multisensory temporal order judgements: the role of hemispheric redundancy, Int. J. Psychophys. 50 (2003) 165–180.
[30] R.J. Zatorre, Neural specializations for tonal processing, Ann. N. Y. Acad. Sci. 930 (2001) 193–210.
[31] R.J. Zatorre, P. Belin, V.B. Penhune, Structure and function of auditory cortex: music and speech, Trends Cogn. Sci. 6 (2002) 37–46.
[32] K. Munhall, E. Vatikiotis-Bateson, Spatial and temporal constraints on audiovisual speech perception, in: G.A. Calvert, C. Spence, B.E. Stein (Eds.), The Handbook of Multisensory Processing, MIT Press, Cambridge, 2004, pp. 177–188.