Audiovisual Synchrony Perception for Complex Stimuli: How "Special" is Speech?

Argiro Vatakis & Charles Spence

Crossmodal Research Laboratory, University of Oxford, Oxford, UK (argiro.vatakis@psy.ox.ac.uk)

Introduction

❖ The majority of studies on the multisensory perception of synchrony have used simple transitory stimuli (such as brief sound bursts and light flashes; e.g., Hirsh & Sherrick, 1961; Zampini, Shore, & Spence, 2003).

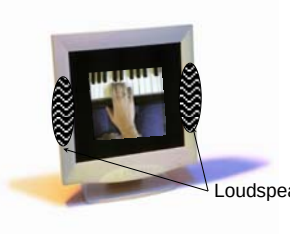
❖ Studies investigating the perception of synchrony with ecologically valid stimuli have typically used continuous audiovisual speech stimuli (Dixon & Spitz, 1980; Grant, van Wassenhove, & Poeppel, 2004).

❖ However, methodological problems (spatial confound, pitch-shifting, and/or criterion shifting) with these studies mean that they might not provide an accurate measure of people's sensitivity to (i.e., ability to discriminate) desynchronized speech.

❖ Consequently, robust evidence regarding people's sensitivity to asynchrony in speech is still lacking. Furthermore, the question of how sensitivity to asynchrony in speech compares to other kinds of complex non-speech stimuli, such as musical instruments remains unanswered. In order to understand speech processing, and its potentially 'special' nature (Massaro, 2004), one needs to compare the perception of synchrony in speech to other complex non-speech events. Here, we examined people's sensitivity to audiovisual asynchrony using both speech and musical stimuli.

Methods

Participants (N = 21) performed a temporal order judgment (TOJ) task using the method of constant stimuli (SOAs = ±400, ±300, ±200, ±100, & 0 ms).



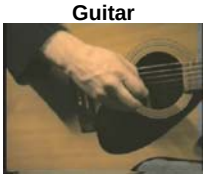
Auditory and visual stimuli presented from the same spatial location

The audiovisual stimuli consisted of twelve video clips presented on a black background, consisting of:

Speech

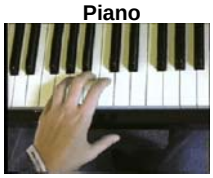
British male saying /a/, /p/, /lo/, or /me/

Guitar

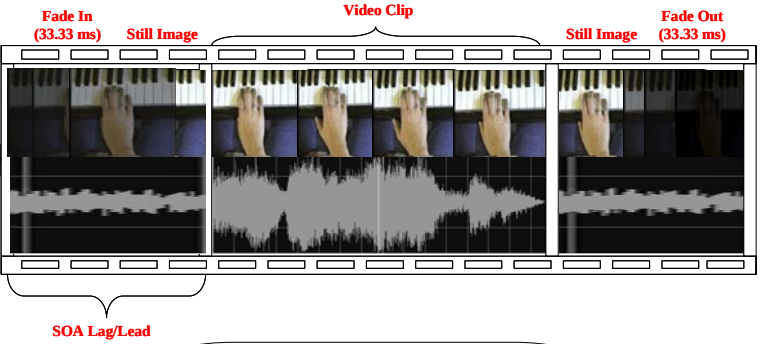


Musical notes "a", "d", "db", or "eg" played on a classical guitar

Piano

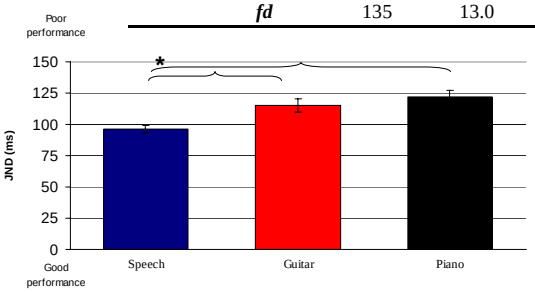


Musical notes "a", "d", "ce", or "fd" played on a piano



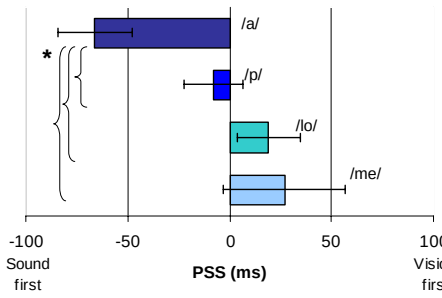
Results

Condition	Category exemplar	JND MEAN	JND SE	PSS MEAN	PSS SE
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



Better temporal discrimination performance (i.e., lower Just Noticeable Difference; JND) for speech stimuli than for guitar or piano music stimuli [F(2,251) = 8.33, p < .01].

Speech stimulus /a/ required an audio lead for the Point of Subjective Simultaneity (PSS) to be achieved. The PSS for /a/ was significantly different from the audio lead required for /p/, or the visual leads required for /lo/ or /me/ [F(3,83) = 7.69, p < .01].



Discussion

❖ These results provide the first empirical findings regarding people's sensitivity to asynchrony in musical stimuli (using brief video clips and the TOJ task).

❖ Our results show that people are better able to detect asynchrony in brief speech videos than in music video clips.

❖ The PSS was shown to depend on the speech sound presented.

❖ JNDs for speech stimuli were higher than those observed in previous studies that have used simple sound-light pairs as experimental stimuli (e.g., Hirsh & Sherrick, 1961; Zampini et al., 2003), but were much smaller than those observed in previous studies of asynchrony detection using continuous speech stimuli (Dixon & Spitz, 1980; Grant et al., 2004). This may be due to:

1) The use of brief stimuli with the controlled viewing of the area around speakers' mouth, & minimal head movements (Thomas & Jordan, 2004);
2) The TOJ task offering a more sensitive index of people's sensitivity to asynchrony than the tasks used in previous studies;
3) The fact that continuous speech consists of a series of words each having a different PSS hence potentially requiring a broad window for multisensory integration.

❖ We are currently investigating the 'unity assumption' by assessing the consequences of mismatching visual lip-movements and auditory speech signals on the multisensory aspects of temporal perception.

References

Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception*, **9**, 719-721.

Grant, K. W., van Wassenhove, V., & Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. *Journal of the Acoustical Society of America*, **108**, 1197-1208.

Hirsh, I. J., & Sherrick, Jr., C. E. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology*, **62**, 424-432.

Massaro, D. W. (2004). From multisensory integration to talking heads and language learning. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processing* (pp. 153-176). Cambridge, MA: MIT Press.

Thomas, S. M., & Jordan, T. R. (2004). Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, **30**, 873-888.

Zampini, M., Shore, D. I., & Spence, C. (2003). Multisensory temporal order judgements: The role of hemispheric redundancy. *International Journal of Psychophysiology*, **50**, 165-180.