Visual lip information supports auditory word segmentation

Antje Strauss, Christophe Savariaux, Sonia Kandel, Jean-Luc Schwartz

Visual lip information supports auditory word segmentation

Antje Strauß, Christophe Savariaux, Sonia Kandel, Jean-Luc Schwartz

CNRS, GIPSA-Lab, F-38000 Grenoble, France; Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France

antje.strauss@gipsa-lab.fr

Introduction

- Speech segmentation has been shown to depend on statistical learning of auditory regularities, e.g. transitional probabilities between syllables [1] and prosodic cues including fluctuations in intensity, F0, segment durations and various articulatory components [2,3].
- However, up to now benefits of visual prosodic cues for speech segmentation have only been investigated in artificial languages [4,5].
- We hypothesize that lip information are used in natural speech when word segmentation is difficult as for example in the case of liaisons in French.

Methods

- **17 French sentences** were created consisting of the carrier phrase "C'est" [engl. "That is"] followed by a determiner and a noun that allowed two possible readings either with liaison (e.g., "l'amarre" [the rope] (A_1)) or without liaison (e.g., "la mare" [the pond] (A_2)).

1 Stimulus design

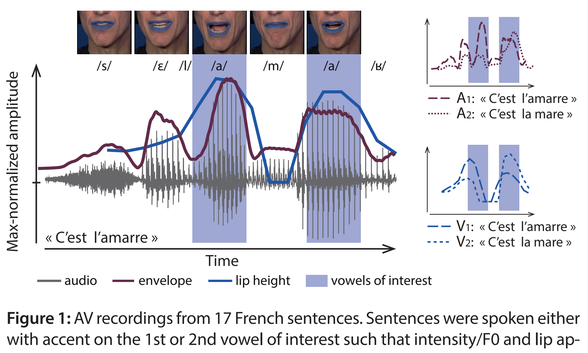


Figure 1: AV recordings from 17 French sentences. Sentences were spoken either with accent on the 1st or 2nd vowel of interest such that intensity/F0 and lip aperture are higher on the emphasized compared to unemphasized vowels.

- The speaker was instructed to produce **10 auditorily hyper-articulated repetitions** of each possible reading (A_1, A_2) and **10 ambiguous** utterances (A_{AM}).
- Afterwards, the speaker was listening to his recordings while producing lip movements in synchrony: **5 times visually hyper-articulating** each possible reading (V_1,V_2).

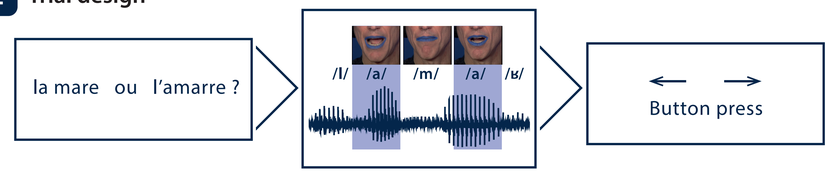
	A1 /l'amarre/	**A2** /la mare/	**AAM** /lamarre/
VNO	A1VNO	A2VNO	AAMVNO
V1 /l'amarre/	A1V1	A2V1	AAMV1
V2 /la mare/	A1V2	A2V2	AAMV2

- Audiovisual stimuli were recorded using a PAL camera (SONY HDR-XR500E) with a sampling rate of **25 images per sec** and an AKG (C-100S) microphone for the audio track.
- **Lips were colored in blue** to be able to apply a chromakey on each image leaving only lip contours. Lip parameters (width, height, surface) were extracted by using the Tacle software developed at Gipsa [6].

Methods

- **21 French native speakers** (normal-hearing; normal or corrected to normal vision; 23.5 ±3.6 years, M±SD) participated in a 2 Alternative Forced Choice task (left, right; counterbalanced across participants).

2 Trial design



Results

3 Stimulus characteristics

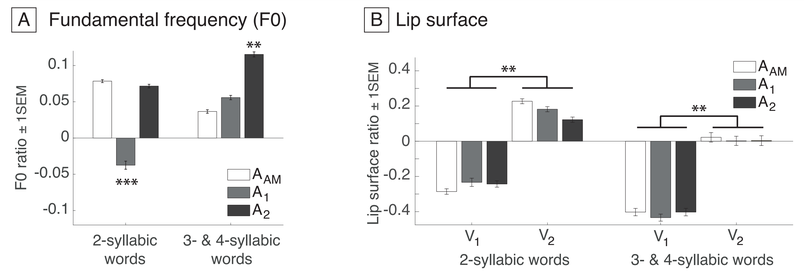


Figure 3: Analysis of stimuli characterisitcs separated for 2- and 3- or 4-syllabic words. A ratio between the two vowels of interest (difference divided by sum) is calculated. Positive values indicate that F0 (A) or lip surface (B) are higher in the second compared to the first vowel of interest. **A.** F0 dissociates A_1 and A_2 but not ambiguous stimuli. **B.** Lip surface dissociates V_1 and V_2.

4 Behavioural results

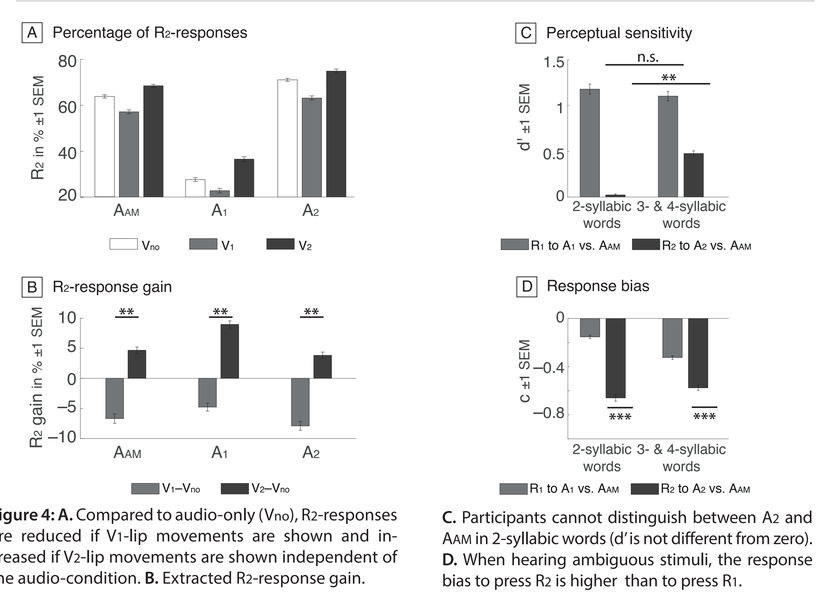


Figure 4: A. Compared to audio-only (V_{no}), R_2-responses are reduced if V_1-lip movements are shown and increased if V_2-lip movements are shown independent of the audio-condition. **B.** Extracted R_2-response gain. **C.** Participants cannot distinguish between A_2 and A_{AM} in 2-syllabic words (d' is not different from zero). **D.** When hearing ambiguous stimuli, the response bias to press R_2 is higher than to press R_1.

Discussion

The current data suggest that visual lip information could have an impact on word segmentation processes. This is particularly relevant for ambiguous utterances like the ones that embed liaisons in French.

1. Lip movements hyper-articulating the CV-segmentation (V_2, e.g. "la mare", in opposition to the VC-segmentation V_1, e.g. "l'amarre") increased decisions for the CV-segmentation in all audiovisual conditions compared to audio-only.
2. Contrary to our prediction, this response gain by lip information was the same even if acoustic cues were ambiguous.

Our data are in line with studies that show an influence of visual lip information when listening to speech.
The results extend evidence showing the usage of lip information in contrastive prosody [7], during multi-stable speech perception [8], and to trigger lexical access [9].
Thus, visual speech does not only provide segmental but also suprasegmental and prosodic cues which enable the perceiver to successfully segment words from a continuous speech stream.

Outlook

5 Is visual hyper-articulation mandatory?

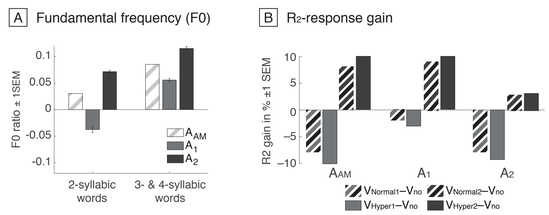


Figure 5. A. Revision of the ambiguous condition such that F0 of A_{AM} really falls in-between A_1 and A_2. R_2-responses should be around 50% **B.** Is the usage of visual prosodic cues weaker if lip movements are not hyper- but only normally articulating? Is the AV-gain only present when information are congruent?

6 EEG: Does theta phase predict segmentation?

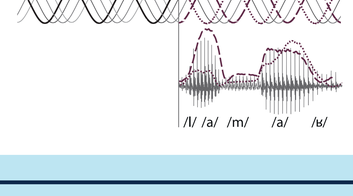


Figure 6. Theta oscillations entrain to speech. Do they align with the most important parts of the speech signal, i.e. word onset, and thus predict the segmentation decision?

References

[1] Cunillera et al. (2010) *Q J Exp Psychol* 63(2):260–74.

[2] Fougeron (2001) *J Phonetics* 29:109–135.

[3] Spinelli et al. (2010) *Atten Percept Psychophys* 72 (3): 775–787.

[4] Sell & Kaschak (2009) *Mem Cognit* 37(6):889–894.

[5] Mitchel & Weiss (2014) *J Neurosci* 32(35):12268–76.

[6] Lallouache (1990) *Proceedings XVIIIèmes Journées d'études sur la Parole*, 282–286.

[7] Dohen et al. (2004) *Speech Communication* 44 (1–4): 155–172.

[8] Sato et al. (2007) *Percept Psychophys* 69(8):1360–72.

[9] Fort et al. (2013) *Lang Cogn Process* 28(8):1207–23.

This research is funded by the European Research Council under the European Community's Seventh Framework Programme (FP7/ 2007-2013 Grant Agreement no. 339152, "Speech Unit(e)s").