Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 513–524
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Speech face perception is locked to anticipation in speech production ☆

Emilie Troille [a,b], Marie-Agnès Cathiard [a,*], Christian Abry [a,1]

[a] CRI, EA 610, Université Stendhal, BP 25, 38040 Grenoble Cedex 9, France
[b] GIPSA-Lab-ICP, UMR 5216, CNRS-INPG-Université Stendhal, BP 25, 38040 Grenoble Cedex 9, France

Received 10 April 2009; received in revised form 30 October 2009; accepted 3 December 2009

Marie and Christian dedicate this article to the late Christian Benoit (–1998) and Tahar Lallouache (–2006), their first companions on speech lips.

Abstract

At the beginning of the 90's, it was definitively demonstrated that as early as the visual speech information is perceivable, speech identification can be processed. Cathiard et al. [Cathiard, M.-A., Tiberghien, G., Tseva, A., Lallouache, M.-T., Escudier, P., 1991. Visual perception of anticipatory rounding during acoustic pauses: a cross-language study. In: Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-en-Provence, France, 4, pp. 50–53] used different V-to-V anticipatory spans, with articulatory measurements, along silent pauses, in a perceptual gating paradigm, and established that up to 200 ms "speech can be seen before it is heard". These results were later framed into the framework of a general anticipatory control model, the *Movement Expansion Model* [Abry, C., Lallouache, M.-T., Cathiard, M.-A., 1996]. How can coarticulation models account for speech sensitivity to audio–visual desynchronization? In: Stork, D., Hennecke, M. (Eds.), Speechreading by Humans and Machines, NATO ASI Series F: Computer, Vol. 150. Springer-Verlag, Berlin, Tokyo, pp. 247–255]. Surprisingly the timing of the vowel and consonant auditory and visual streams remained until now poorly understood within the typical CVCV span. A first preliminary test was published by Escudier, Benoît and Lallouache [Escudier, P., Benoît, C., Lallouache, M.-T., 1990. Identification visuelle de stimuli associés à l'opposition /i/-/y/: étude statique. Colloque de physique, supplément au n° 2, tome 51, 1er Congrès Français d'Acoustique, C2-541-544]: this is the issue we took up again more than 10 years later. And for the first time we found that "speech can be heard before it is seen". The main purpose of the present contribution will be to bring new data in order to clear up apparent contradictions, essentially due to misconceptions of variability and lawfulness in speakers' behavior.

Keywords: Auditory-visual speech perception; Speech production; Anticipation

1. Introduction

At the beginning of the 90s, it was definitively demonstrated that as early as the visual speech information is perceivable, speech identification can be processed. Cathiard et al. (1991; see also Cathiard, 1994; Cathiard et al., 1996) used different V-to-V anticipatory spans, with articulatory measurements, along silent pauses, in a perceptual gating paradigm, and found that up to 200 ms "speech can be seen before it is heard". The available lip image processing system (Lallouache, 1991) allowed to evidence that speech face perception was fairly locked to anticipation in production. Accordingly these results could be framed in the framework of a general anticipatory control model,

☆ A part of this contribution, a preliminary analysis of Experiment 1, has been presented at AVSP'07: Troille, E., Cathiard, M.-A., Abry, C., 2007. Consequences on bimodal perception of the timing of the consonant and vowel audiovisual flows. In: Proceedings of the International Conference on Audio–Visual Speech Processing, 31 août-3 Septembre, Hilvarenbeek, Pays-Bas, pp. 281–286.

* Corresponding author. Address: 10 Chemin du Ruy F-38690 Chabons, France. Tel.: +33 (0)4 76 65 08 25.
E-mail addresses: emilie.troille@gmail.com (E. Troille), marieagnes.cathiard@u-grenoble3.fr (M.-A. Cathiard), chris.abry@orange.fr (C. Abry).
[1] Present address: 10 Chemin du Ruy F-38690 Chabons, France.

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.12.005

the *Movement Expansion Model* (its presentation atop of the preceding models was made available 10 years ago by Farnetani in the reference book on *Coarticulation...*, Hardcastle, 1999; for the link with perception, see Abry et al. (1996); for production testing of the model, since Abry and Lallouache (1995a,b), for French, see recently Noiray et al. (2006, 2008), for French children and English).

It is important to note that the more classical CVCV phonetic span remained until now poorly understood as concerns the audiovisual production-perception phenomenology, more specifically, the coordination of the vowel and consonant bimodal streams. Smeele et al. (1994, Smeele, 1994) in a pure perceptual gating experiment (with no articulatory measurements) assessed, plainly, that the most visible places for consonants, bilabials and labiodentals, were naturally better identified when vision was added, just along two 40 ms steps in the constriction phase before the release.

To our knowledge the first preliminary attempt to measure the time course of audio and visual perception in CVCVs, with the control tracking of lip gestures, was published by Escudier et al. (1990).

The test stimulus chosen after many exchanges with Christian Abry was [zizy], i.e.: (i) with a vowel-to-vowel rounding gesture, (ii) throughout a fricative voiced consonant, this (iia) in order to interrupt as less as possible the acoustic flow, (iib) while offering a sufficiently high frequency frication noise, which could carry the resonance changes corresponding to the lip rounding gesture, above the range of the formants characterizing the change in the vowels.

They found roughly (apart from methodological problems they acknowledged) that the visible rounding anticipatory gesture was perceived 40–60 ms ahead of the acoustic change, what confirmed the lead of vision on audition. In this exemplar, the rounding gesture was largely anticipated since it began in the previous [i] vowel and it was yet achieved at the beginning of the consonant. The acoustic analysis consisted of the follow-up of vocalic formants only; so no information was given about the contribution of the frication noise of the consonant. The third formant downward movement began into the [i] vowel; but it could just be perceived into the first part of the [z] consonant, probably due to the downward movement of resonances into the high frequency frication noise produced by the rounding gesture. So the [y] vowel was seen early into the [i] vowel, but could only be heard into the [z] consonant.

This is the paradigm we took up again more than 10 years later. And for the first time in this over studied research domain we found that "speech could be heard before it is seen" (Troille et al., 2007). A new performance of a [zizy] stimulus was recorded, produced by the same male speaker as in Escudier et al.'s study. Then audio, visual and audiovisual perceptual tests were designed. The real new thing was that auditory perception was 45 ms ahead of visual perception. But since, this time, besides the visual tracking of lip area and the auditory tracking of formants, we could study the downward movement of resonances along the frication noise of the [z] consonant, we will be able to account below for this result, at first glance a surprising one (see Section 2).

In this contribution, we added again to this paradigm, recording this [zizy] stimulus produced by a female speaker (see Section 3). Contrary to the preceding results, the perceptual tests evidenced no difference between the perception of the three auditory, visual and audiovisual modalities. Hence in this third case, "speech can be heard as early as it is seen". By analysing the stimulus, we will account why, in this case, there was clearly no superior modality.

The main purpose of the present contribution will be to clear up what could appear as contradictory results. They will essentially be attributed to current misconceptions of variability and lawfulness in speakers' behavior. A variability which perceivers have to cope with, since we demonstrated that with their ears and/or their eyes, they succeed in recovering in due time the recoverable linguistic information in the modality they have prior access to.

2. Experiment 1

The purpose of this experiment was initially to test if, given the three possible VCV stimulus internal structure, regarding the AV timing, speech was seen before being heard. We will see that the acoustic and optical properties are crucial to lock perceptual behavior.

2.1. Recording

A French male speaker (the same as in Escudier et al.'s experiment) was audiovisually recorded. He produced in random order 12 repetitions of the two sequences "T'as dit ZIZU ze?" and "T'as dit ZIZI ze?" ("Did you say...?"). These sentences were chosen to explore the whole transition, from the vowel [i] to the vowel [y], with an intervocalic [z] consonant. It is well-known indeed that this fricative is permeated by audible vocalic coarticulation effects (Whalen, 1984). The video recording was carried out in an anechoic booth, at 50 frames/s, with front and profile views, with the sound sampled at 22.05 kHz. Blue make-up applied to the lips of the speaker allowed an automatic incrustation of a saturated black, using a Chroma-Key, in order to perform an accurate detection of labial parameters, notably small between-lip areas. The *ICP-Lip-Shape-Tracker* (Lallouache, 1991) via automatic image processing of the videos, provided a set of different lip parameters every 20 ms (frame by frame).

2.2. Data

Two articulatory parameters which characterize fairly well the rounding gesture, i.e. upper lip-protrusion (P1)