Audio Visual Integration with Competing Sources in the Framework of Audio Visual Speech Scene Analysis

Attigodu Chandrashekara Ganesh, Frédéric Berthommier, Jean-Luc Schwartz

Audio Visual Integration with Competing Sources in the Framework of Audio Visual Speech Scene Analysis

Attigodu Chandrashekara Ganesh, Frédéric Berthommier and Jean-Luc Schwartz

Abstract We introduce "Audio-Visual Speech Scene Analysis" (AVSSA) as an extension of the two-stage Auditory Scene Analysis model towards audiovisual scenes made of mixtures of speakers. AVSSA assumes that a coherence index between the auditory and the visual input is computed prior to audiovisual fusion, enabling to determine whether the sensory inputs should be bound together. Previous experiments on the modulation of the McGurk effect by audiovisual coherent vs. incoherent contexts presented before the McGurk target have provided experimental evidence supporting AVSSA. Indeed, incoherent contexts appear to decrease the McGurk effect, suggesting that they produce lower audiovisual coherence hence less audiovisual fusion. The present experiments extend the AVSSA paradigm by creating contexts made of competing audiovisual sources and measuring their effect on McGurk targets. The competing audiovisual sources have respectively a high and a low audiovisual coherence (that is, large vs. small audiovisual comodulations in time). The first experiment involves contexts made of two auditory sources and one video source associated to either the first or the second audio source. It appears that the McGurk effect is smaller after the context made of the visual source associated to the auditory source with less audiovisual coherence. In the second experiment with the same stimuli, the participants are asked to attend to either one or the other source. The data show that the modulation of fusion depends on the attentional focus. Altogether, these two experiments shed light on audiovisual binding, the AVSSA process and the role of attention.

Keywords Audio visual binding · Auditory speech analysis · McGurk effect · Attention

A. C. Ganesh (✉) · F. Berthommier · J.-L. Schwartz
Grenoble Images Parole Signal Automatique-Lab, Speech and Cognition Department, CNRS, Grenoble University, UMR 5216, Grenoble, France
e-mail: ganesh.attigodu@gipsa-lab.grenoble-inp.fr

F. Berthommier
e-mail: frederic.berthommier@gipsa-lab.grenoble-inp.fr

J.-L. Schwartz
e-mail: jean-luc.schwartz@gipsa-lab.grenoble-inp.fr

P. van Dijk et al. (eds.), *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing,* Advances in Experimental Medicine and Biology 894,
DOI 10.1007/978-3-319-25474-6_42

1 Introduction

This paper is focused on a tentative fusion between two separate concepts: Auditory Scene Analysis and Audio-Visual fusion in speech perception.

Auditory Scene Analysis (ASA) introduced the principle of a two-stage process in the auditory processing of complex auditory scenes with competing sources (Bregman 1990). A first stage would involve segmenting the scene into auditory elements, which would be segregated or grouped in respect to their common source, either by bottom-up innate primitives or by learnt top-down schemas. Decision and formation of a final percept would be done at a second later stage.

Audio-Visual fusion in speech perception refers to the well-known fact that speech perception involves and integrates auditory and visual cues, as shown in various paradigms such as speech in noise (Sumby and Pollack 1954; Erber 1969) or the perception of conflicting stimuli (the so-called McGurk effect, McGurk and MacDonald 1976; also see Tiippana 2014).

Since a pioneer proposal by Berthommier (2004), our group proposed that auditory scene analysis and multisensory interactions in speech perception should be combined into a single "Audio-Visual Speech Scene Analysis" (AVSSA) process. The basic claim is that the two-stage analysis-and-decision process at work in ASA should be extended to audiovisual speech scenes made of mixtures of auditory and visual speech sources. A first audiovisual binding stage would involve segmenting the scene into audiovisual elements, which should be segregated or grouped in respect to their common multisensory speech source, either by bottom-up *audiovisual primitives* or by learnt top-down *audiovisual schemas*. This audiovisual binding stage would control the output of the later decision stage, and hence intervene on the output of the speech-in-noise or McGurk paradigms.

To provide evidence for this "binding and fusion" AVSSA process, Nahorna et al. (2012, 2015) showed that the McGurk effect can be significantly and strongly reduced by an audiovisual context made of a few seconds of incoherent material (sounds and images coming from different speech sources) presented before the McGurk target (audio "ba" plus video "ga"): the target, classically perceived as "da" in the McGurk effect, was more often perceived as "ba", suggesting a decreased weight of the visual input in the fusion process. The interpretation was that the incoherent context resulted in an "unbinding" effect decreasing the visual weight and hence diminishing the McGurk effect. This modulation of the McGurk effect through incoherent contexts was further extended to speech in noise (Ganesh et al. 2013), and a possible neurophysiological correlate of the binding/unbinding process was provided in an EEG experiment (Ganesh et al. 2014).

However, these studies were based on audiovisual scenes that never implied competing sources. The objective of the present study was to test the "binding and fusion" AVSSA process in scenes including competition between audiovisual sources. For this aim, we generated two audiovisual sources, one made of a sequence of isolated syllables, and the other one made of a sequence of sentences. We prepared two kinds of combinations, with the same auditory content (mixing the two audio sources, syllables and sentences) and two different video contents,

either the syllables or the sentences. These two combinations ("Video syllables" and "Video sentences") were used as the context in a McGurk experiment. We hypothesized that since syllables correspond to stronger audiovisual modulations in time and hence stronger audiovisual coherence than sentences, the association between the visual input and the corresponding auditory input would be stronger for syllables than for sentences. Hence the coherence of the audiovisual context would be stronger for syllables, and it would lead to a larger visual weight and more McGurk effect than with visual sentences. Furthermore, the introduction of a competition between sources made it possible to introduce attention factors in the paradigm, and we tested whether the attentional focus put by the participants on either syllables or sentences would play a role in the AVSSA process.

2 Method

2.1 Participants

The study involved twenty-nine French participants without hearing or vision problems (22 women and 7 men; 27 right-handed and 2 left handed; mean age = 29.2 years; SD = 10. 4 years), who all gave informed consent to participate in the experiment.

2.2 Stimuli

The stimuli were similar to those of the previous experiment by (Nahorna et al. 2015) with suitable modification in the paradigm. They were prepared from two sets of audiovisual material, a "syllables" material and a "sentences" material, produced by a French male speaker, with lips painted in blue to allow precise video analysis of lip movements (Lallouache 1990). The whole experiment consisted of two types of contexts followed by a target.

2.2.1 Target

The target was either a congruent audiovisual "ba" syllable ("ba-target" in the following), serving as a control—or an incongruent McGurk stimulus with an audio "ba" mounted on a video "ga" ("McGurk target" in the following).

2.2.2 Context

There were two types of contexts i.e. "Video syllables" and "Video sentences". In both contexts, the set of audio stimuli was the same. It consisted of a sequence of 2

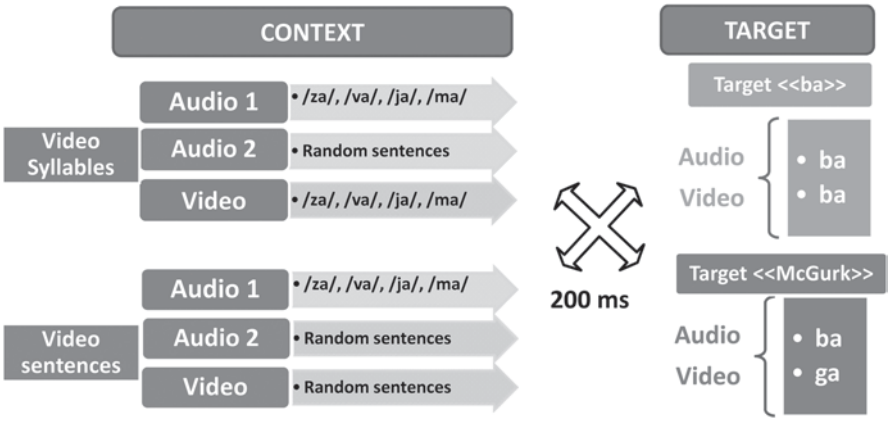

Fig. 1 Description of the audiovisual material

or 4 syllables (A-syl-2 or A-syl-4) randomly extracted from the set "pa," "ta," "va," "fa," "za," "sa," "ka," "ra," "la," "ja," "cha," "ma," or "na," mixed with a portion of a random set of sentences with the adequate duration (A-sent-2 or A-sent-4). The 2- vs. 4-syllable duration was selected from earlier experiments by Nahorna et al. (2015), showing that the effect of incoherent context was maximal (maximal reduction of the McGurk effect) for short 2-syllable contexts and slightly less for longer 4-syllable contexts. The visual components of the context were the visual stream associated with either the auditory syllables (V-syl-2 or V-syl-4) or the auditory sentences (V-sent-2 or V-sent-4). Therefore, in the "Video syllables" contexts, there was an audiovisual "syllables" source competing with an audio "sentences" source, while in the "Video sentences" contexts, there was an audiovisual "sentences" source competing with an audio "syllables" source (Fig. 1). A 200 ms fading transition stimulus (five images) was implemented between context and target to ensure continuity between images.

There were altogether 120 stimuli with four times more "McGurk" than "Ba" targets (serving as controls), and with the same number of occurrences of the V-syl-2, V-syl-4, V-sent-2 and V-sent-4 contexts (6 occurrences each for "Ba" targets, 24 occurrences each for McGurk targets). Exactly the same set of 30 targets was presented after the 4 types of contexts. The 120 stimuli were presented in a random order and concatenated into a single 7-min film.

2.3 Procedure

The study included two consecutive experiments, Exp. A. followed by Exp. B (always in this order). In Exp. A, the participants were involved in a monitoring paradigm in which they were asked to constantly look at the screen and monitor for possible "ba" or "da" targets by pressing an appropriate key, as in Nahorna et al. (2012, 2015). In Exp. B the monitoring "ba" vs. "da" task remained the same (with a different order of the 120 stimuli in the film), but in addition, specific instructions

were given to participants, either to put more attention to syllables ("Attention syllables") or to put more attention to sentences ("Attention sentences"). The order of the "Attention syllables" and "Attention sentences" conditions was counterbalanced between the participants. To increase the efficiency of the attentional demand, participants were informed that they would be questioned on the content of either the "syllables" or the "sentences" material at the end of the experiment. A practice session was provided for all of them and most of the participants were indeed able to recall specific syllables or words. The films were presented on a computer monitor with high-fidelity headphones set at a comfortable fixed level.

2.4 Processing of Responses

Response time was computed in reference to the acoustic onset of the burst of the "b" in the target syllable, discarding values higher than 1200 ms or lower than 200 ms. "ba" and "da" responses were taken into account only when they occurred within this time window (200–1200 ms) and in case of two different responses inside the time window, both responses were also discarded. Finally, for each participant and each condition of context and target (and attention in Exp. B), a global score of "ba" responses was calculated as the percentage of "ba" responses divided by the sum of "ba" and "da" responses to the target, and a mean response time was calculated as the average of response times for all the responses to the target.

3 Results

First of all, the mean percentage of "ba" scores for McGurk targets over all conditions in Exp. A was computed for each subject, and participants providing mean scores larger than 95 % or less than 5 % were discarded, considering that these subjects provided either too strong or too low McGurk effects to enable binding modulations to be displayed. This resulted in discarding 8 out of 29 participants. All further analyses for both Exp. A and B will hence concern only the 21 remaining subjects. As expected, the global score (percentage of "ba" responses relative to "ba" + "da" responses) for all control "ba" targets was close to 100 % in all conditions in both experiments. Therefore, from now on we will concentrate on McGurk targets.

3.1 On the Role of Context Type Without Explicit Attention Focus (Exp. A)

Percentages of "ba" responses to McGurk targets in Exp. A (without explicit attentional focus) are displayed on the left part of Fig. 2. A two-factor repeated measures ANOVA with *context type* ("Video syllables" vs. "Video sentences") and *context*

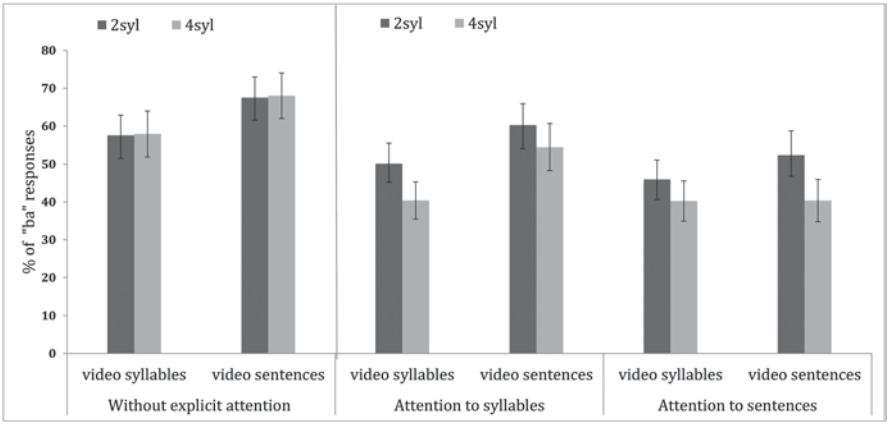

Fig. 2 The percentage of "ba" responses (relative to the total number of "ba" or "da" responses) for "McGurk" targets, in the "Video syllables" vs. "Video sentences" contexts in Experiment A and Experiment B

duration (2- vs. 4-syllables) as the independent variables was administered on these percentages (applying Greenhouse-Geisser correction when applicable). The effect of *context type* is significant [$F(1, 20) = 34.65$, $p < 0.001$], with a higher McGurk effect (10% less "ba" responses) with the "Video syllables" context. This is in line with our prediction that audiovisual coherence is higher in the "Video syllables" condition, leading to a higher binding level, a larger visual weight and hence a larger number of McGurk fusion ("da" responses). *Context duration* displayed no significant effect on "ba" scores, either in isolation or in interaction with *context type*.

3.2 On the Interaction Between Context Type and Attention Focus (Exp. B)

Percentages of "ba" responses to McGurk targets in Exp. B (involving explicit attention towards one or the other source) are displayed on the right part of Fig. 2. A repeated-measures ANOVA was administered on these percentages with three factors, *context type* ("Video syllables" vs. "Video sentences"), *context duration* (2- vs 4-syllables) and *attention* ("Attention syllables" vs. "Attention sentences") by applying Greenhouse-Geisser correction when applicable.

The effect of context type [$F(1, 20) = 11.91$, $p < 0.001$] is significant, as in Exp. A: video syllables produce more McGurk than video sentences. Contrary to Exp. A, the effect of context duration [$F(1, 20) = 33.86$, $p < 0.001$] is also significant, with no interaction with context type. The attention factor alone is not significant, but its interaction with context type is significant [$F(1, 20) = 11.07$, $p < 0.005$]. Post-hoc analyses with Bonferroni corrections show that while there is no significant difference between the two attention conditions for the "Video syllables" context type,

there is a difference for the "Video sentences" condition, with a lower "ba" percentage (a higher McGurk effect) in the "Attention sentence" condition. Interestingly, while the "ba" percentage is higher for the "Video sentences" than for the "Video syllables" condition when attention is put in syllables, there is no more significant difference when attention is put on sentences.

Finally, the three-way interaction between context type, context duration and attention is significant [$F(1, 20) = 6.51$, $p < 0.05$], with a larger difference between durations from the "Video syllables" to the "Video sentences" condition in the "Attention sentences" than in the "Attention syllables" condition.

3.3 Response Time

The results are consistent with the previous findings (Nahorna et al. 2012) in which response times were larger for McGurk targets, independently on context. In both experiments and in all contexts, the processing of "ba" responses was indeed quicker compared to McGurk responses. Two-way repeated-measures ANOVA on *target* and *context type* in Exp. A displays an effect of target [70 ms quicker response for "ba" targets, $F(1, 20) = 14.25$, $p < 0.005$] and no effect of context or any interaction effect. A three-way repeated-measures ANOVA on *condition type*, *attention* and *target* in Exp. B displays once again an effect of target [80 ms quicker response for "ba" targets, $F(1, 20) = 4.47$, $p < 0.05$] and no other significant effect of other factors, alone or in interaction.

4 Discussion

Audiovisual fusion has long been considered as an automatic fusion process (e.g. Massaro 1987). However Exp. A. confirms that the contextual stimulus may modulate fusion as in our previous experiments (Nahorna et al. 2012; Ganesh et al. 2013; Nahorna et al. 2015) and extends the concept to the case of competing sources.

This supports the AVSSA hypothesis, in which a first speech scene analysis process would group together the adequate audiovisual pieces of information and estimate the degree of audiovisual coherence. The effect of context type (larger McGurk effect in the "Video syllables" condition) could be due to the differences in audiovisual correlations for syllables and sentences. Indeed, correlation analysis between audio (full band envelope) and video (mouth opening area) material for syllables and sentences provides a mean correlation value of 0.59 for "Video syllables" and 0.10 for "Video sentences" (Fig. 3). Another factor could increase binding with syllables, i.e. the presence of a streaming mechanism in which the syllabic target would be associated to the syllables stream rather than to the sentences stream.

A number of recent experiments pointed the role of general attentional mechanisms able to globally decrease the amount of fusion (Tiippana et al. 2004; Alsius

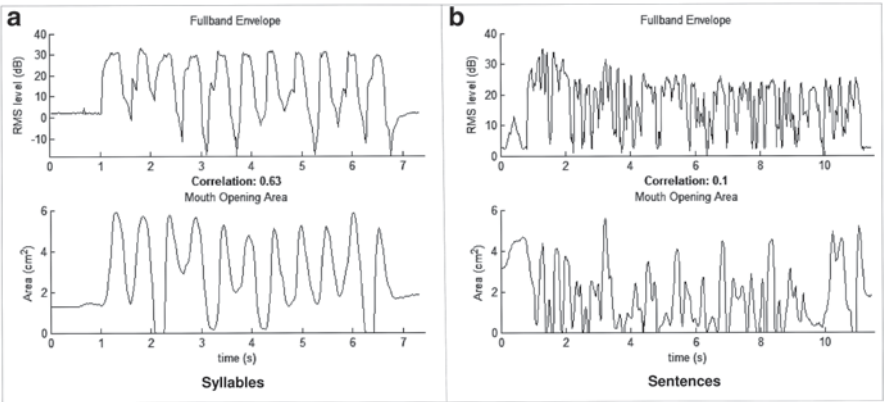

Fig. 3 Variations in time of the audio full band envelope (*top row*) and the video mouth opening area (*bottom row*) for syllables (**a**, *left*) and sentences (**b**, *right*). Notice that the fluctuations in time of the audio and video information are much more coherent between the audio and the video streams for syllables than for sentences

et al. 2007; Navarra et al. 2010; Alsius et al. 2014). Experiment B shows that attentional mechanisms may intervene at the level of single audiovisual sources in an audiovisual speech scene, selectively increasing or decreasing the amount of fusion depending on the coherence of the attended source. Interestingly, attention intervened only for "Video sentences". Our interpretation is that binding could be pre-attentive for syllables, because of their salient audiovisual comodulations making them pop out as strong bottom-up audiovisual primitives. In contrast, since the coherence of AV sentences is low, the attentional focus could enhance audiovisual top-down schemas increasing binding.

These two studies provide confirmation and development to the view that audiovisual fusion in speech perception includes a first stage of audiovisual speech scene analysis. A number of previous studies suggested that the presentation of a visual stream can enhance segregation by affecting primary auditory streaming (Rahne et al. 2007; Marozeau et al. 2010; Devergie et al. 2011) or that visual cues can improve speech detection and cue extraction (Grant and Seitz 2000; Kim and Davis 2004; Schwartz et al. 2004; Alsius and Munhall 2013); though some contradictory studies highlight cases where unimodal perceptual grouping precedes multisensory integration (Sanabria et al. 2005).

Altogether, the "binding stage" in the AVSSA process, in which the coherence between auditory and visual features would be evaluated in a complex scene, provides a mechanism in order to properly associate the adequate components inside a coherent audiovisual speech source. The present study confirms the importance of this mechanism in which the "binding stage" enables the listener to integrate "similar sources" or segregate "dissimilar sources" in Audio Visual fusion.

Acknowledgments This project has been supported by Academic Research Community "Quality of life and ageing" (ARC 2) of the Rhône-Alpes Region, which provided a doctoral funding for Ganesh Attigodu Chandrashekara.

References

Alsius A, Munhall KG (2013) Detection of audiovisual speech correspondences without visual awareness. Psychol Sci 24(4):423–431. doi:10.1177/0956797612457378

Alsius A, Navarra J, Soto-Faraco S (2007) Attention to touch weakens audiovisual speech integration. Exp Brain Res 183(3):399–404. doi:10.1007/s00221-007-1110-1

Alsius A, Mottonen R, Sams ME, Soto-Faraco S, Tiippana K (2014) Effect of attentional load on audiovisual speech perception: evidence from ERPs. Front Psychol 5:727. doi:10.3389/fpsyg.2014.00727

Berthommier F (2004) A phonetically neutral model of the low-level audio-visual interaction. Speech Comm 44(1–4):31–41. doi:10.1016/j.specom.2004.10.003

Bregman AS (1990) Auditory scene analysis. MIT Press, Cambridge

Devergie A, Grimault N, Gaudrain E, Healy EW, Berthommier F (2011) The effect of lip-reading on primary stream segregation. J Acoust Soc Am 130(1):283–291. doi:10.1121/1.3592223

Erber NP (1969) Interaction of audition and vision in the recognition of oral speech stimuli. J Speech Hear Res 12(2):423–425

Ganesh AC, Berthommier F, Schwartz J-L (2013). Effect of context, rebinding and noise on audiovisual speech fusion. Paper presented at the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), Lyon, France

Ganesh AC, Berthommier F, Vilain C, Sato M, Schwartz J-L (2014) A possible neurophysiological correlate of audiovisual binding and unbinding in speech perception. Front Psychol 5:1340. doi:10.3389/fpsyg.2014.01340

Grant KW, Seitz PF (2000) The use of visible speech cues for improving auditory detection of spoken sentences. J Acoust Soc Am 108(3):1197–1208. doi:10.1121/1.1288668

Kim J, Davis C (2004) Investigating the audio–visual speech detection advantage. Speech Comm 44(1–4):19–30. doi:10.1016/j.specom.2004.09.008

Lallouache MT (1990). Un poste 'visage-parole.' Acquisition et traitement de contours labiaux (A 'face-speech' workstation. Acquisition and processing of labial contours). Paper presented at the Proceedings of the eighteenth Journées d'Etudes sur la Parole, Montréal, QC

Marozeau J, Innes-Brown H, Grayden DB, Burkitt AN, Blamey PJ (2010) The effect of visual cues on auditory stream segregation in musicians and non-musicians. PLoS ONE 5(6):e11297. doi:10.1371/journal.pone.0011297

Massaro DW (1987) Speech perception by ear and eye. Lawrence Erlbaum Associates, Hillsdale

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264(5588):746–748 [10.1038/264746a0]

Nahorna O, Berthommier F, Schwartz JL (2012) Binding and unbinding the auditory and visual streams in the McGurk effect. J Acoust Soc Am 132(2):1061–1077. doi:10.1121/1.4728187

Nahorna O, Berthommier F, Schwartz JL (2015) Audio-visual speech scene analysis: characterization of the dynamics of unbinding and rebinding the McGurk effect. J Acoust Soc Am 137(1):362–377. doi:10.1121/1.4904536

Navarra J, Alsius A, Soto-Faraco S, Spence C (2010) Assessing the role of attention in the audiovisual integration of speech. Inf Fusion 11(1):4–11. doi:10.1016/j.inffus.2009.04.001

Rahne T, Bockmann M, von Specht H, Sussman ES (2007) Visual cues can modulate integration and segregation of objects in auditory scene analysis. Brain Res 1144:127–135. doi:10.1016/j.brainres.2007.01.074

Sanabria D, Soto-Faraco S, Chan J, Spence C (2005) Intramodal perceptual grouping modulates multisensory integration: evidence from the crossmodal dynamic capture task. Neurosci Lett 377(1):59–64. doi:10.1016/j.neulet.2004.11.069

Schwartz JL, Berthommier F, Savariaux C (2004) Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition 93(2):B69–B78. doi:10.1016/j.cognition.2004.01.006

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26(2):212–215. doi:10.1121/1.1907309

Tiippana K (2014) What is the McGurk effect? [Opinion]. Front Psychol 5. doi:10.3389/fpsyg.2014.00725

Tiippana K, Andersen TS, Sams M (2004) Visual attention modulates audiovisual speech perception. Eur J Cog Psychol 16(3):457–472. doi:10.1080/09541440340000268