Journal of Phonetics 49 (2015) 41–54

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Research Article

Exploring the contribution of prosody and gesture to the perception of focus using an animated agent☆

Pilar Prieto[a,b,*], Cecilia Puglesi[a], Joan Borràs-Comes[a], Ernesto Arroyo[a], Josep Blat[a]

[a] *Universitat Pompeu Fabra, Barcelona, Spain*
[b] *Institució Catalana de Recerca i Estudis Avançats (ICREA) – Universitat Pompeu Fabra, Departament de Traducció i Ciències del Llenguatge, Despatx 53.720, Campus de la Comunicació – Poblenou, C/Roc Boronat, 138, 08018 Barcelona, Spain*

ARTICLE INFO

Article history:
Received 10 June 2013
Received in revised form
10 October 2014
Accepted 16 October 2014

Keywords:
Contrastive focus
Intonation
Prosody
Gestures
Eyebrow movements
Head movements
Talking heads
Audiovisual prosody
Multisensory perception

ABSTRACT

Speech prosody has traditionally been analyzed in terms of acoustic features. Although visual features have been shown to enhance linguistic processing, the conventional view is that facial and body gesture information in oral (non-signed) languages tends to be redundant and has the role of helping the hearer recover the meaning of an utterance. Though prosodic information in face-to-face communication is produced with concurrent visual information, little is known about their audiovisual multisensory interactions. We conducted two perception experiments modeled after the McGurk paradigm with a 3D animated character, in which varying degrees of discordance between auditory and visual information were created to investigate two related questions regarding the detection of contrastive-corrective focus: (a) how important are gestural cues with respect to auditory cues and (b) what is the relevance of the different gestural movements involved (i.e., head nodding, eyebrow raising)? Participants were presented with combinations of auditory and visual cues for both information and Contrastive Focus Statements (Experiment 1, with the corresponding unimodal control experiments) or combinations of two visual cues (namely combinations of competing eyebrow and head movements) without auditory information (Experiment 2), and were asked to identify whether the utterance presented was a statement or a correction. Results of Experiment 1 showed that (a) the presence of either acoustic or gestural features of contrastive focus were key in guiding the listener towards one interpretation or another, and (b) listeners were more sensitive to one of the modalities when the other was weaker. Results of Experiment 2 showed that (a) both types of visual cues (head and eyebrow movements) contributed individually to the perception of contrastive focus, and (b) head nods were more informative than eyebrow movements for focus identification. Overall, our findings suggest that prosodic and visual information work in a complementary fashion and are not integrated in the same way as auditory and visual information during segmental perception.

1. Introduction

The study of speech prosody has traditionally been based on acoustic information related to f0 pitch movements (i.e., information about the intonation contour of an utterance) and information related to the duration and intensity of the sound chain. Yet in face-to-face communication prosodic features are typically produced with correlated visual features, such as head and eyebrow movements, or body and arm/hand movements, which are temporally synchronized with prominent prosodic units and which can be very helpful in the processing of speech. Classic experiments have shown that visual cues contribute to speech intelligibility and the detection of segmental information in noisy conditions (Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Sumby & Pollack, 1954),

☆A preliminary version of this article was presented at *Interspeech* 2011 (Florence, August 27–31). We are grateful to the audience at that conference, and especially to David House, for very helpful comments. We are particularly indebted to Aurora Bel for her help in recruiting participants for the experimental sessions and to Javier Agenjo for his assistance with the NINOs platform. This research was supported by the following grants: BFU2012-31995 and TIN2008-05995 Learn3 CICYT (awarded by the Spanish Ministry of Science and Innovation), 2014 SGR 925 (Generalitat de Catalunya), and iMP (European Union 7th Framework project). The second author also acknowledges a MAEC-AECID Grant (Spanish Ministry of Foreign Affairs and Cooperation).

* Corresponding author at: Institució Catalana de Recerca i Estudis Avançats (ICREA) – Universitat Pompeu Fabra, Departament de Traducció i Ciències del Llenguatge, Despatx 53.720, Campus de la Comunicació – Poblenou, C/Roc Boronat, 138, 08018 Barcelona, Spain. Tel.: +34 93 225 48 99; fax: +34 93 542 16 17.

E-mail addresses: pilar.prieto@upf.edu (P. Prieto), ceciliapuglesi@gmail.com (C. Puglesi), joan.borras@upf.edu (J. Borràs-Comes), ernesto.arroyo@upf.edu (E. Arroyo), josep.blat@upf.edu (J. Blat).

URL: http://prosodia.upf.edu/home/en/index.php (P. Prieto).

0095-4470/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2014.10.005

with the study by Munhall et al. (2004) focusing on head movement in particular. They used an animated character to assess speech-in-noise comprehension in the presence of normal head movements, exaggerated head movements, or no head movements, and found that performance comprehension increased in the presence of natural head movements. Al Moubayed and Beskow (2009) also showed that word recognition and intelligibility using animated talking heads increased when focally-accented (i.e., prominent) words were supplemented with head-nods or eyebrow raising gestures.

Prominence is one of the prosodic functions that have been studied in most detail, and it has been shown to be strongly correlated with the presence of certain gestural features. A body of research has reported that in natural communication prominent words tend to be accompanied by head nods and eyebrow movements or by more exaggerated movements of the articulators (Dohen, 2009; Dohen, Lœvenbruck & Hill, 2006; Ekman, 1979; Graf, Cosatto, Strom & Huang, 2002; Swerts & Krahmer, 2008), and these gestures also act as conversational signals (e.g., raised eyebrows to express interest on the part of the listener or nodding to provide conversational encouragement; Ekman, 1979). Dohen et al. (2006) tracked the movement of speakers' faces while producing prosodic contrastive focus in French and assessed head, eyebrow, and cheek movements as well as the opening, spreading, and protrusion of the lips, and chin movements. They found that contrastive focus in French is produced using lengthening and over-articulation of the focused constituent. With regard to articulatory correlates, lip protrusion had the largest correlation with the presence of stress, which suggests that this cue may be of particular importance, at least in French. They also found that focus was sometimes signaled by raised eyebrows and/or head nodding, but the link was highly inter- and intra-speaker dependent. Production studies have also reported on the fine temporal alignment that exists between the most prominent part of head gestures (the apex) and accented syllables in speech (i.e., prosodically prominent syllables) (Borràs-Comes, Vanrell, and Prieto, 2014; Loehr, 2007, and others).

In the perceptual domain, several studies have demonstrated the usefulness of eyebrow and head movements to facilitate the audiovisual perception of speech prominence (House, Beskow, & Granström, 2001 for Swedish; Krahmer, Ruttkay, Swerts, & Wesselink, 2002a and Krahmer & Swerts, 2006 for Dutch; Massaro & Beskow, 2002 for English; Dohen & Lœvenbruck, 2009 for French). Many of these studies have manipulated the presence vs. absence of prosodic and visual information. For example, Swerts and Krahmer (2008) showed that conflicting auditory prosodic and visual information (specifically, head nod, eyebrow raising, and manual beat gestures, as well as pitch range and duration values) affect the perceived location of emphatic words in a phrase. In their study, participants listened to recordings of a sentence with three prosodically prominent words whose auditory and visual prominence cues were manipulated. These cues were either congruent (i.e., occurring on the same word) or incongruent (i.e., occurring in such a fashion that the auditory and visual cues were positioned on different words). Their results showed that participants could more easily determine prominence when the visual cue occurred on the same word as the auditory cue, while displaced visual cues hindered prominence perception. Dohen and Lœvenbruck (2009) found that prosodic contrastive focus could be easily detected on the basis of either the audio or the visual modality alone (when the visual modality included only the lip contour information), leading to a ceiling effect. In an experiment involving whispered speech, they reported that a combination of auditory and visual information constituted an advantage for the perception of prosodic features. Specifically, the study showed that adding vision to audition improved focus detection and also reduced reaction times.

Thus, though most of these studies suggest that gestural dynamics do convey important information that may improve the perception of focus in conversational situations, little is known about the extent to which different activations of visual cues to prominence interact with different activations of prosodic information in the perception of speech. Previous work on the integration of speech information with visual cues from a speaker's face has concentrated on the segmental effects, i.e., the recognition of vowels and consonants. McGurk and MacDonald's (1976) classic study showed that presentation of speech information from the voice with incompatible concurrent presentation of gestural information from the face leads to confusion and illusory percepts. In some cases, subjects reported hearing sounds that were not provided either by the voice alone or by the movements of the face alone (in one of the most dramatic cases, when an auditory “baba” was combined with a visual “gaga”, some speakers reporting hearing “dada”). The McGurk effect thus shows that speech perception of the segmental content does not simply superimpose audition on or add audition to vision, but rather that the two modalities are integrated. Yet speech is not unique in being perceived by both ear and eye. Work on the detection of emotions (deGelder and Vroomen, 2000) has also shown clear interactions between vision and audition, that is, when a fearful voice is juxtaposed against a happy face, the voice is perceived to be happier. Similarly, while linguistic prominence is also produced in a multimodal fashion, little is known about how visual cues influence the perception of prosodic categories. To help fill this gap in the research, in the present study we use a novel method of testing multimodal perception of contrastive focus that is comparable to methods used in experiments that examine the multimodal identification of phonetic segments. The first aim of the study is to evaluate the relevance of prosodic and visual information in the perception of contrastive focus in an experiment which presented participants with combinations of different activations of prosodic and gestural features. Experiment 1 examines the potential influence of the presence of gestural markers (specifically, head and eyebrow movements) on the perception of contrastive focus when they are competing with auditory information. By using controlled manipulations of intonation (i.e., pitch range variation) and gestural variation through animated agents (i.e., variation in the strength of head nod and eyebrow raising movements) we were able to obtain data on the relative contributions of intonation and gesture to the perception of contrastive focus.

Recent work has also sought to determine which visual features are most effective in perceptual terms for conveying prominence and focus in speech. Though results are partially contradictory, it seems that visual cues from the upper face together with head movements are powerful cues to prominence when synchronized with the stressed vowel of the prominent word. Swerts and Krahmer (2008) investigated which area of a speaker's face contains the strongest cues to prominence, using stimuli with either the entire face visible or only parts of it. Results showed that the upper facial area (i.e., eyebrow movements) had a stronger cue value for