Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 57 (2014) 317–330

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Tracking eyebrows and head gestures associated with spoken prosody

Jeesun Kim [a], Erin Cvejic [a,b], Chris Davis [a,*]

[a] The MARCS Institute, University of Western Sydney, Australia
[b] School of Psychiatry, University of New South Wales, Australia

Available online 15 June 2013

Abstract

Although it is clear that eyebrow and head movements are in some way associated with spoken prosody, the precise form of this association is unclear. To examine this, eyebrow and head movements were recorded from six talkers producing 30 sentences (with two repetitions) in three prosodic conditions (Broad focus, Narrow focus and Echoic question) in a face to face dialogue exchange task. Movement displacement and peak velocity were measured for the prosodically marked constituents (critical region) as well as for the preceding and following regions. The amount of eyebrow movement in the Narrow focus and Echoic question conditions tended to be larger at the beginning of an utterance (in the pre-critical and critical regions) than at the end (in the post-critical region). Head rotation (nodding) tended to occur later, being maximal in the critical region and still occurring often in the post-critical one. For eyebrow movements, peak velocity tended to distinguish the regions better than the displacement measure. The extent to which eyebrow and head movements co-occurred was also examined. Compared to broad focussed condition, both movement types occurred more often in the narrow focussed and echoic question ones. When these double movements occurred in narrow focused utterances, brow raises tended to begin before the onset of the critical constituent and reach a peak displacement at the time of the critical constituent, whereas rigid pitch movements tended to begin at the time of critical constituent and reach peak displacement after this region. The pattern for echoic questions was similar for eyebrow motion however head rotations tended to begin earlier compared to the narrow focus condition. These results are discussed in terms of the differences these types of visual cues may have in production and perception.

Keywords: Visual prosody; Eyebrow movements; Focus; Sentence modality; Guided principal component analysis

1. Introduction

In addition to the movements of the lips, mouth and jaw that are inherent to speech production, a speaker will also move her/his head and eyebrows. How do these peri-oral movements relate to how speech is uttered, i.e., speech prosody? The answer to this question has implications for areas such as character animation and the development of auditory–visual speech synthesis (e.g., for interactive conversational agents), and also bears on theories concerning the relationship between speech planning and speech accompanying gestures (see Kendon, 2004).

The bulk of the research that has investigated the association between visual speech (movements of the talker's head and face likely to be visible to an interlocutor) and properties of the auditory speech signal has examined oral motion (i.e., movements of the lips, mouth and jaw). Results have generally shown that there are consistent across-talker visual speech correlates for prosodic features like contrastive focus, although different talkers were found to have different pre- and post-focal articulation strategies (Dohen and Lœvenbruck, 2009; Dohen et al., 2009). Finding that there is an association between oral motion and speech acoustics is perhaps unsurprising. That is, since many properties of auditory and visual speech originate from the same spatiotemporal event (i.e., speech production), it might be expected that articulatory (i.e., lip, mouth and jaw opening) and other related movements (e.g., cheek

[*] Corresponding author. Address: The MARCS Institute, University of Western Sydney, Locked Bag 1797, PENRITH, NSW 2751, Australia. Tel.: +61 2 9772 6855; fax: +61 2 9772 6040.
E-mail address: chris.davis@uws.edu.au (C. Davis).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.06.003

motion) would have a close relationship with acoustic speech properties in general (Yehia et al., 1998), and with those used to signal prosody more specifically. For instance, in order to produce a speech sound over an extended duration (a property commonly reported for narrowly focused and echoically questioned syllables), the speaker must maintain the configuration of the articulators for this amount of time (de Jong, 1995). Similarly, increases in amplitude (also associated with prosodically marking an important or questioned constituent) are likely to be accompanied by more dynamic jaw movements that end in a lower jaw position (Edwards et al., 1991; van Summers, 1987).

Studies of the relationship that peri-oral movements have with speech acoustics have generally examined auditory properties identified as cues for prosody (e.g., intensity, *F*0). For example, Hadar and colleagues (Hadar et al., 1983) found a relationship between intensity modulation of speech and rigid head movement, and several studies have shown that eyebrow movements are associated with the modulation of *F*0 (Cavé et al., 1996; Guaïtella et al., 2009). It has also been reported in Yehia et al. (2002) that there is a strong association between *F*0 modulation and rigid head motion. Here, it was found that a large amount of variance in *F*0 (88% for an American English speaker, and 73% for a speaker of Japanese) could be estimated from rigid head motion. Although these correlations were high, the relationship between head motion and *F*0 varied from utterance to utterance, and for some tokens the correlation between head motion and *F*0 was very low. This variability in the coupling between *F*0 and head motion suggests that the association may reflect particular speaker communicative strategies; something that may vary according to whether the speaker can see the interlocutor or not (Cvejic et al., 2012; Fitzpatrick et al., 2011).

Studies of the relationship between gestures such as eyebrow raises and rigid head motion and speech acoustics that signal linguistic prosodic contrasts have tended to use relatively unconstrained speech production procedures. For example, Flecha-García (2010) audio-visually recorded pairs of participants who engaged in task-oriented face-to-face dialogues (i.e., a map task). The auditory and visual recordings were then annotated offline for the occurrence of pitch accents in the auditory signal (by a single coder) and for eyebrow raises (defined as any upward movement from a neutral baseline position of at least one eyebrow) in the visual signal (by two coders). Swerts and Krahmer (2010) examined the eyebrow movements and head nods of newsreaders to determine how these were associated with the relative prominence of spoken words (visual feature labelling was conducted by two coders who labelled rapid eyebrow movements and head nods).

The results of these studies indicated that these peri-oral movements were aligned with pitch accents. For instance, in Flecha-García (2010), more than 80% of eyebrow movements started within 330 ms of the nearest pitch accent, with the average eyebrow raise occurring 60 ms before the onset of a pitch accented syllable. In the Swerts and Krahmer, (2010)'s study of four Dutch newsreaders, it appeared that the degree to which head and eyebrow movements aligned with auditory focus (pitch accent) was conditioned by the strength of the accent. That is, 70% of strong pitch accents (items identified as emphasised by more than half of the coders) were accompanied by an eyebrow raise and 89% were by a rigid head movement. In contrast, weakly accented words (those identified by less than half of the coders) were accompanied by head movements on only 40% of occasions, and similarly (37%) by eyebrow movements. It is worth noting that eyebrow movements were coded for almost a quarter of non-accented words (23.2%), whereas few rigid head movements were coded for these items. These results were taken as evidence that talkers align the occurrence of non-articulatory visual prosodic cues with auditory correlates of prosody in order to maximise the strength of the prosodic contrast conveyed to the perceiver.

The above results suggested that both rigid head movement (nods) and eyebrow movements may be co-ordinated with auditory cues for prosodic contrasts; however more data is required in order to establish the reliability and consistency of these findings. Indeed, Flecha-García (2010) herself calls for data from a larger number of speakers (she used three) and from different types of speech production setting. Moreover, new collection methods may be required, since the frame-rates of video-based motion analysis potentially limit the resolution of the temporal alignment of the auditory and motion data and video-based labelling makes the collection of accurate and consistent data difficult. New procedures are also needed in order to overcome inconsistencies in how many suitable prosodic contrasts are produced so that a relatively balanced amount of data can be gathered across participants and conditions.

The current study meets some of these requirements by examining the association between the acoustic properties associated with two types of prosodic contrasts (corrective focus and sentence modality) and comparing these to a broad focus condition. These different prosodic types were used to examine how head and eyebrow motions were affected by the generation of a narrow focus compared with forming an echoic question (a change in sentence modality). Six speakers producing 30 sentences with different prosody were recorded using a constrained experimental setup in which motion data were obtained by motion capture. Here, the temporal distribution of peri-oral gestures and their relationship with the prosodically marked constituents can be accurately determined (at 60 Hz) along with measures of motion velocity that would be difficult to obtain using video mark-up analysis.

Before describing the data collection setup in detail, it is important to consider how to determine if a rigid head movement or eyebrow movement occurred. For eyebrow movements, Flecha-García (2010) used a criterion in which only eyebrow raises were counted. Cavé et al. (1996) stipu-