Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 57 (2014) 233–243

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Analysis of relationship between head motion events and speech in dialogue conversations

Carlos Toshinori Ishi [a,*], Hiroshi Ishiguro [b], Norihiro Hagita [a]

[a] *ATR Intelligent Robotics and Communication Labs., Kyoto, Japan*
[b] *ATR Hiroshi Ishiguro Special Lab., Kyoto, Japan*

Available online 15 June 2013

Abstract

Head motion naturally occurs in synchrony with speech and may convey paralinguistic information (such as intentions, attitudes and emotions) in dialogue communication. With the aim of verifying the relationship between head motion events and speech utterances, analyses were conducted on motion-captured data of multiple speakers during spontaneous dialogue conversations. The relationship between head motion events and dialogue acts was firstly analyzed. Among the head motion types, nods occurred with most frequency during speech utterances, not only for expressing dialogue acts of agreement or affirmation, but also appearing at the end of phrases with strong boundaries (including both turn-keeping and giving dialogue act functions). Head shakes usually appeared for expressing negation, while head tilts appeared mostly in interjections expressing denial, and in phrases with weak boundaries, where the speaker is thinking or did not finish uttering. The synchronization of head motion events and speech was also analyzed with focus on the timing of nods relative to the last syllable of a phrase. Results showed that nods were highly synchronized with the center portion of backchannels, while it was more synchronized with the end portion of the last syllable in phrases with strong boundaries. Speaker variability analyses indicated that the inter-personal relationship with the interlocutor is one factor influencing the frequency of head motion events. It was found that the frequency of nods was lower for dialogue partners with close relationship (such as family members), where speakers do not have to express careful attitudes. On the other hand, the frequency of nods (especially of multiple nods) clearly increased when the inter-personal relationship between the dialogue partners was distant.

Keywords: Head motion; Paralinguistic information; Dialogue act; Inter-personal relationship; Spontaneous speech

1. Introduction

Head motion naturally occurs in synchrony with speech utterances, and may carry paralinguistic information related to intentions, attitudes or emotion in dialogue communication. Therefore, a better understanding of the relationship between head motion events and speech utterances is important for application in multi-modal human–agent or human–robot interactions.

Head motion analyses can be focused on two problems from the application viewpoint: one is how to generate the head motion of CG (Computer Graphics) agents or robots, synchronized with their speech utterances (e.g. Yehia et al., 2002; Munhall et al., 2004; Watanabe et al., 2004; Sargin et al., 2006; Beskow et al., 2006; Busso et al., 2007; Foster and Oberlander, 2007; Hofer and Shimodaira, 2007); the other is how to recognize the user's head motion and interpret its role in communication (e.g., Iwano et al., 1996; Watanuki et al., 2000; Graf et al., 2002; Dohen et al., 2006; Sidner et al., 2006; Morency et al., 2007; Burnham et al., 2007). The generation of a natural head motion is not only useful for improving human–agent or human–robot interaction, but can also improve intelligibility in noisy environments. For example, it is reported that a

[*] Corresponding author. Address: ATR Intelligent Robotics and Communication Lab, 2-2-2 Hikaridai, "Keihanna Science City", Kyoto 619-0288, Japan. Tel.: +81 774 95 2457; fax: +81 774 95 1408.
E-mail address: carlos@atr.jp (C.T. Ishi).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.06.008

better perception of syllables has been achieved in a speech-in-noise task, with the normal, natural head motion compared with speech without head motion and only auditory stimulus, in an experiment with animations (Munhall et al., 2004). Intelligibility of tones may also be improved by use of head motion in tonal languages (Burnham et al., 2007).

There are many works in the literature, which analyzed the correlation between head motion and prosodic features, such as the fundamental frequency (F0) contours (which represent pitch movements) (Yehia et al., 2002; Munhall et al., 2004; Busso et al., 2007).

For example, Yehia et al. tried to associate head motion with speech over the fundamental frequency (F0) (Yehia et al., 2002). Experiments using read speech utterances of one American English speaker (ES) and one Japanese speaker (JS) showed the following results for estimation of head motion from F0 and vice versa. From head motion to F0, the average correlation was 0.73 for JS and 0.88 for ES. Opposite estimation from F0 to head motion showed a less obvious correlation (0.25 for JS and 0.50 for ES, on average). In addition, correlation among F0 and the 6DOF (degrees-of-freedom) (3DOF for rotation and 3DOF for translation) of head motion was between 0.39 and 0.52 for ES, and between 0.22 and 0.30 for JS, which are in average less than 0.50. Munhall et al. (2004) reports that head motion is correlated with pitch and amplitude of the talker's voice, in Japanese read speech utterances, regarding all 6 DOF. For several sentences, correlations were almost always over 0.50, on average about 0.63.

The results above imply that the correspondence between head motion and prosodic features is language dependent, and that the use of only prosodic information might not be enough to generate natural head motion.

Other works show that head motion may differ according to emotional states (Beskow et al., 2006; Busso et al., 2007). Analysis of the relation between facial parameters (including head motion) and several expressive modes was also reported by Beskow et al. for short read Swedish utterances in which focal accent was systematically varied (Beskow et al., 2006). Results indicated that in all expressive modes, words with focal accent were accompanied by a greater variation of the facial parameters than the words in non-focal positions. Regarding head motion, it is reported that head pitch has larger variations in "certain", "angry" and "confirming" modes, head yaw in "angry" and "happy" modes, and head roll in "certain" modes. Busso et al. compared head motion of neutral and emotional speech for synthesis purposes (Busso et al., 2007). They investigated head motion for four emotional states (neutral, sadness, happiness and anger). As prosodic features, they used the pitch (F0), the RMS (root mean square) energy and their 1st and 2nd derivatives. For the head motion they took account of the 3 DOF of head rotation. Canonical correlation analysis (which provides a measure of the correlation between two streams of data with equal or different dimensionality) was applied for the streams of prosodic features and head motion, resulting in correlations around 0.7 for all expressive modes. However, as the prosodic features are implicit in the synthesis models, it is not clear which of the features are related to a specific emotion.

Most of the works focusing on head motion synthesis, as in the ones cited above, usually associate the acoustic features directly with the raw head motion measurements through models. However, to better understand the roles of head motion in speech communication, it could be more appropriate to analyze head motion events, like nods, head shakes and head tilts.

For example, head nods are reported to be related to emphasis (or focus) (Sargin et al., 2006; Foster and Oberlander, 2007; Dohen et al., 2006). Variations in speech emphasis depending on head movement were observed by Graf et al., for English sentences (Foster and Oberlander, 2007). They reported that emphasis of a word often goes along with head nodding, and a rise of the head can correspond with a rise in the voice. They call these movements 'visual prosody'. A talking head which included 'visual prosody' through head motion was reported as looking more natural even if this motion was not really connected with the content of the spoken text. Sargin et al. also reported correlation between head motion events (nods and head tilts) and speech prominences marked as pitch accents for English (Sargin et al., 2006). They carried out experiments with one native speaker of Canadian English to investigate the correlation between keyword speech (like "left", "right" and "straight") and gestures (including hand and head gestures). With focus on head nods and tilts, a correspondence of about 64% between these two head motion types and pitch accents was reported. Dohen et al. also reported that eyebrow raising and/or head nods signal focus in French (Dohen et al., 2006).

As can be observed from the past works described above, most of them focus on the relationship between head motion and prosodic features. However, we consider that this relationship might be language-dependent, since the function of the prosodic features differs, for example, if the language is a tonal language (such as Chinese and Thai), a lexical pitch-accent language (such as Japanese), or a stress-accent language (such as English and other European languages). The present work focuses on Japanese, whose correlation between head motion and prosodic features has been reported to be lower than in English (Yehia et al., 2002). Further, it can also be observed that most of the past works described above analyzed read speech or acted emotional speech data for few speakers. Thus, the results reported in the past works may not be applied for any language.

In the present work, the analysis of the relationship between head motion and speech are focused on Japanese spontaneous speech. For Japanese, there are works reporting that head motion might be related to turn-taking and speech act functions, for spontaneous dialogue speech. For example, Iwano et al. analyzed relations between head motion and the semantics of utterances in Japanese spoken