Research Report

Visual Prosody and Speech Intelligibility

Head Movement Improves Auditory Speech Perception

K.G. Munhall,[1,2,3] Jeffery A. Jones,[3] Daniel E. Callan,[3] Takaaki Kuratate,[3] and Eric Vatikiotis-Bateson[3,4]

[1]Department of Psychology and [2]Department of Otolaryngology, Queen's University, Kingston, Ontario, Canada; [3]ATR-International Human Information Science Laboratories, Kyoto, Japan; and [4]Department of Linguistics, University of British Columbia, Vancouver, British Columbia, Canada

ABSTRACT—*People naturally move their heads when they speak, and our study shows that this rhythmic head motion conveys linguistic information. Three-dimensional head and face motion and the acoustics of a talker producing Japanese sentences were recorded and analyzed. The head movement correlated strongly with the pitch (fundamental frequency) and amplitude of the talker's voice. In a perception study, Japanese subjects viewed realistic talking-head animations based on these movement recordings in a speech-in-noise task. The animations allowed the head motion to be manipulated without changing other characteristics of the visual or acoustic speech. Subjects correctly identified more syllables when natural head motion was present in the animation than when it was eliminated or distorted. These results suggest that nonverbal gestures such as head movements play a more direct role in the perception of speech than previously known.*

During natural face-to-face conversations, a wide range of visual information from the movements of the face, head, and hands is available to conversational partners. In the work reported here, we studied the impact on speech perception of watching one component of this rich visual stimulus—a talker's head movements. It is well known that the intelligibility of degraded auditory speech is enhanced when listeners view a talker's lip movements (Sumby & Pollack, 1954). Watching these lip movements can also influence the perception of perfectly audible speech (McGurk & MacDonald, 1976) or be the sole basis of speech perception (Bernstein, Demorest, & Tucker, 2000). The contribution of nonverbal gestures such as head movements[1] to the understanding of spoken language, however, is not well understood.

Previous work on head gestures during speech focused on documenting their timing and motor organization (Hadar, Steiner, Grant, & Rose, 1983, 1984; Hadar, Steiner, & Rose, 1984). These studies suggested that head movements are linked to the production of suprasegmental features of speech such as stress, prominence, and other aspects of prosody. Consistent with the kinematic research of Hadar et al., several studies have shown that subjects can use head and eyebrow movements to determine which word in a sentence is receiving emphatic stress (e.g., Bernstein, Eberhardt, & Demorest, 1998; Risberg & Lubker, 1978; Thompson, 1934) and to discriminate statements from questions (Bernstein et al., 1998; Fisher, 1969; Nicholson, Baum, Cuddy, & Munhall, 2002; Srinivasan & Massaro, 2002).

In this study, we extended this research to test whether visual prosody, as embodied in head motion, plays a role in spoken word recognition. Auditory prosodic structure can aid the segmentation of words from the continuous stream of speech, as well as facilitate lexical access (see Cutler, Dahan, & van Donselaar, 1997, for a review). To examine whether visual prosody functions similarly, we tested whether the presence of visible head motion improved the intelligibility of Japanese sentences in a speech-in-noise task. In order to carry out this test, we created a stimulus set consisting of an animated talking face whose characteristics were initially derived directly from recordings of the face and head motion of a Japanese speaker (Kuratate, Yehia, & Vatikiotis-Bateson, 1998). The advantage of using animation is that head motion can be systematically varied independently of the acoustics and face motion in order to determine the influence of head motion on speech perception.

Address correspondence to K.G. Munhall, Department of Psychology, Queen's University, Kingston, ON, Canada K7L 3N6; e-mail: munhallk@psyc.queensu.ca.

[1]By nonverbal, we mean that the movements are not directly involved in the production of sound. We are not referring to symbolic gestures, such as head nodding to signify "yes" or head shaking to signify "no." The motions that we tested are the arbitrary, rhythmic movements of the head that always accompany speech.

The existence of visual prosody effects would pose a number of challenges for current research on spoken word recognition. For example, the pool of potential cues for segmentation and lexical selection would have to be expanded to the visual modality; acoustic prosody is well documented, but its visual counterpart has been explored in only a preliminary manner. Study of the time course of word activation would have to reconcile differences in processing auditory and visual information at both neural and behavioral levels of observation. Finally, the neural substrates attributed to lexical processing would have to include the areas responsible for visual and audiovisual processing of prosody.

METHOD

Participants

Twelve normal young adults served as subjects. All subjects were native speakers of Japanese with no hearing, speech, or language disorders and normal or corrected-to-normal vision.

Stimuli

Four different audiovisual versions of 20 sentences selected from the Advanced Telecommunications Research Japanese Sentences were produced using a custom animation system (Kuratate et al., 1998). The animations (Fig. 1) were derived from the recorded kinematics of face and head motion produced by a male native speaker of Japanese. The data were collected and analyzed for a speech production study independent of the present research. As part of that study, three-dimensional head and face motion were partitioned through rigid body analysis of the head movements (Horn, 1987). The orientation and position of the head at each instant in time were specified by translation along the three Cartesian axes and rotation about each axis. Once the rigid body coordinates of the head were computed, head motion could be removed from face motion or added in to provide natural or exaggerated amounts of head motion.

The fundamental frequency (F0) and root mean square (RMS) amplitude of the voicing portions of each sentence were computed using Matlab routines (Yehia, Kuratate, & Vatikiotis-Bateson, 2002). Figure 2a shows the voice signal, the resulting F0 track, and the six degrees of freedom of head motion plotted over time. After appropriate down-sampling of the acoustic signals to match the sampling rate of head motion (10 kHz $\rightarrow$ 60 Hz), multiple regression analysis was used on a sentence-by-sentence basis to test the relationship between the head motion and the F0 and RMS-amplitude signals; time-varying features of the acoustics were estimated from the time-varying head motion. Because each sentence contained silent intervals corresponding to pauses and the oral closure intervals of unvoiced consonants (e.g., *p*, *t*, *k)*, only portions of the voice and head-motion signals for each sentence were used in the computation. The sections of the F0, RMS-amplitude, and head-motion signals corresponding to silent intervals were deleted.

Figure 2b shows the percentage of variance (R^2) in F0 accounted for by the motion of the head on a sentence-by-sentence basis. The correlations are consistently high, with on average more than 63% of the variance in the talker's voice pitch being accounted for by the six components of head motion. Figure 2c shows the percentage of variance in the RMS amplitude of the voice accounted for by the motion of the head. Although the correlations are somewhat smaller than for F0 (mean $R^2 = .324$), there are still strong correlations for most sentences. For a more detail description of these relationships, see Yehia et al. (2002).

Fig. 1. Two views of the animated face with different speech sounds being produced and with the head at different positions and orientations.

The head and facial motion in these recordings were manipulated to produce the animated stimuli used in this study. Four versions of each sentence were produced: (a) a version with the recorded natural head motion and the recorded facial motion; (b) a version with zero head motion but the recorded facial motion (the animated face was centered in the screen); (c) a version with double head motion, meaning that the amplitude of head movement was doubled in all six degrees of freedom, while the face articulated the recorded facial motion; and (d) an auditory-only version in which the screen was blackened. Because of the difficulty of animating eye movement correctly and the significance of gaze behavior, the animated face was depicted wearing sunglasses.

Because the recorded acoustics had the same time course as the original face and head motions, synchronization with the animations was trivial. The acoustics were mixed with a commercial multispeaker babble track (Auditec, St. Louis, MO) in order to reduce intelligibility during audiovisual testing. The signal-to-noise ratio was determined by pilot work and held constant for all conditions for all subjects. Specifically, we wanted to ensure low enough intelligibility of the audio signal so that the visual contribution to audiovisual percep-

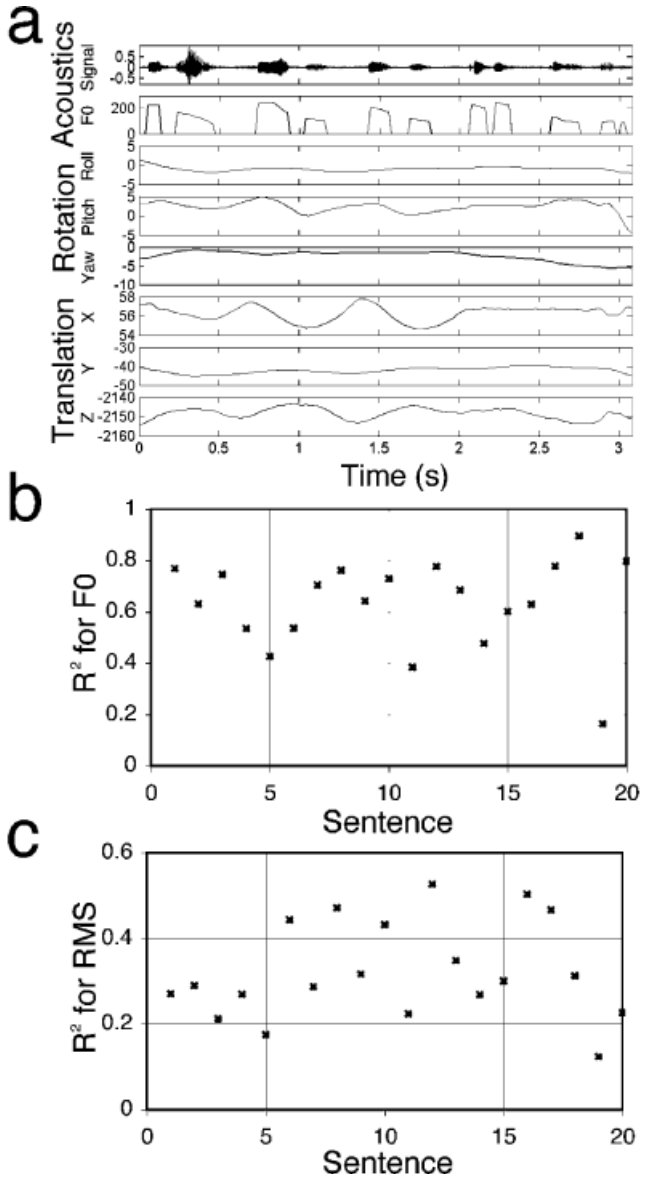

Fig. 2. Example of acoustic and head-movement patterns for a single sentence (a) and results of the multiple regression analyses (b, c). In (a), the acoustic waveform, fundamental frequency of the voice (F0), and six degrees of freedom of head motion (three positional coordinates: x, y, z; three angular coordinates: roll, pitch, yaw) are plotted as a function of time. The R^2 values for the multiple regression analysis of F0 (b) and RMS amplitude (c) as predicted by the six degrees of freedom of head motion are plotted for the 20 sentences used as stimuli.

tion could be reliably assessed. In our experience, 35 to 45% auditory-only intelligibility is sufficiently low to avoid any ceiling effect on the multimodal percept.

Procedure

The intelligibility of the different versions of the sentences was tested in a speech-in-noise task. The subjects' task was to identify as many words as possible. Assignment of sentences to the four conditions was counterbalanced across subjects. The presentation of stimuli was randomized across conditions within subjects. After the presentation of a sentence, the subjects verbally repeated as much of the sentence as they could, and the experimenter recorded their response before initiating the next sentence. Responses were scored in terms of the percentage of Japanese syllables (hiragana) correctly perceived and analyzed using analysis of variance.

RESULTS AND DISCUSSION

When the animated stimuli were presented in noisy listening conditions, the intelligibility varied as a function of the head-motion condition. Figure 3 shows that subjects perceived more hiragana correctly in each of the audiovisual conditions than in the auditory-only condition, indicating that the animation reproduced realistic face motion. Significantly, the best performance occurred with the animation containing natural head motion. These patterns were supported statistically. There was a main effect of condition, $F(3, 33) = 15.65$, $p < .001$. Accuracy in all three audiovisual conditions was reliably better than accuracy in the auditory-only condition (ps $< .001$, .01, and .001 for natural head motion, double head motion, and zero head motion, respectively). In addition, accuracy was greater for the natural-head-motion condition than for both the other audiovisual conditions (ps $< .001$ and .05 for double head motion and zero head motion, respectively), which did not differ statistically ($p > .2$).

The results demonstrate that the stimuli used in this perceptual experiment contained a systematic, and to some extent redundant, relationship between head movement and the acoustic features of prosody. That is, the F0 and amplitude of the voice were highly correlated with the kinematics of head motion during the production of

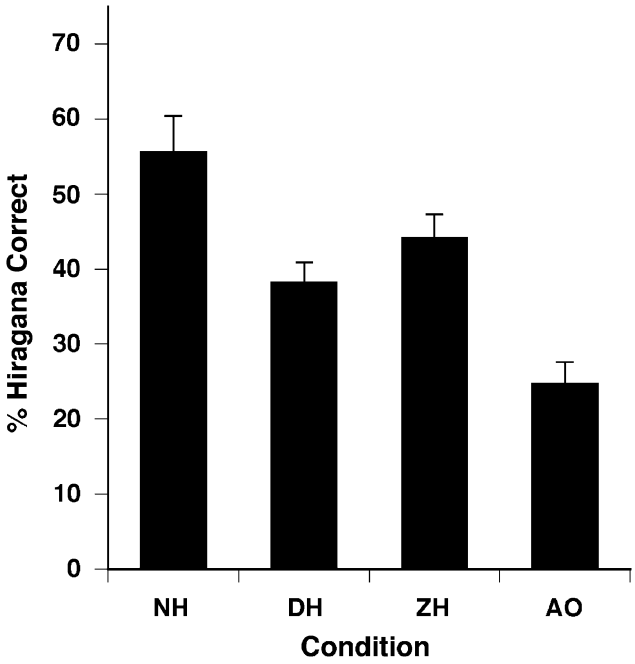

Fig. 3. Percentage of Japanese syllables (hiragana) correctly perceived as a function of animation condition: normal head motion (NH), double the amplitude of head motion in all six degrees of freedom (DH), zero head motion (ZH), and auditory only (AO). Error bars indicate the standard errors of the mean.

the sentences. This audiovisual correlation is consistent with the findings of previous studies of Japanese, Swedish, French, and English speakers (Badin et al., 2002; Graf, Cosatto, Strom, & Huang, 2002; Granström, House, & Lundeberg, 1999; Jiang, Alwan, Keating, Auer, & Bernstein, 2002; Yehia et al., 2002; Yehia, Rubin, & Vatikiotis-Bateson, 1998), which showed strong correspondence between visible movements and acoustic parameters.

The audiovisual correlation measured in the production data was preserved in the animated stimuli created from the measured face- and head-motion data. The differential intelligibility in the different head-motion conditions demonstrates that subjects make use of the visual structure identified by the correlation analysis of the original production data.

Previous evidence for the perception of visual prosody has come primarily from studies in which subjects were asked to differentiate different classes of sentences, such as questions versus statements (Graf et al., 2002; House, 2002; House, Beskow, & Granström, 2001; Srinivasan & Massaro, 2002). These studies showed that head and eyebrow movement are significant cues for the perception of the interrogative/declarative distinction and that the strength of the effect is influenced by the timing of the gestures.

The present study is unique in that it demonstrates an interaction between visual prosody and the identification of individual words in a set of statement sentences. There are several possible mechanisms that could account for the perceptual data reported here. Given the correlations between head movement and the acoustics of the voice, the head gestures may contribute to word processing much the way acoustic prosody is believed to do. Recent evidence suggests that in some languages, acoustic cues for prosody enhance spoken word recognition (Cutler et al., 1997) and may be similar to segmental structure in constraining initial word activation (Soto-Faraco, Sebastián-Gallés, & Cutler, 2001). For example, Japanese subjects can use pitch accent to distinguish ambiguous word fragments (Cutler & Otake, 1999). Alternatively, head motion may act as a timing signal that aids the segmentation of the stream of speech. In this case, enhancements in intelligibility would be related to the shared metrical structure of the auditory and visual signals.

In some ways, it may seem surprising that these visual prosody effects have been demonstrated in Japanese. Studies have shown that Japanese subjects are less susceptible to the McGurk effect than English speakers (e.g., Sekiyama & Tohkura, 1993), and this has led to the view that vision plays a reduced role in Japanese speech perception. However, Japanese speakers show strong audiovisual effects (Massaro, Tsuzaki, Cohen, Gesi, & Heredia, 1993), as well as strong visual effects in noisy listening conditions (Sekiyama & Tohkura, 1991). More significantly, auditory prosody in Japanese has been shown to influence the process of word recognition. Japanese listeners are sensitive to mora boundaries (e.g., Cutler & Otake, 1994) and can use pitch accent information to select word candidates (Cutler & Otake, 1999). Whether the finding reported here extends to other languages needs to be tested.

The head movements that we considered in this study are produced idiosyncratically by talkers (Hill & Johnston, 2001) in rhythm with their speech and by themselves convey no inherent lexical meaning. Their ubiquity, however, suggests that they may be an integral part of communication. The data reported here demonstrate that the movement of the head during speech has functional significance in processing audiovisual speech information. Understanding the linguistic and social intent during a conversation requires processing a wide range of verbal and nonverbal signals. To date, little attention has been paid to the nonverbal context of audiovisual speech. Our data suggest that this aspect of spoken language is not independent from speech perception, and our findings underscore the complex nature of face-to-face communication.

Acknowledgments—This work was funded by grants from the National Institute of Deafness and Other Communicative Disorders (DC05774), Natural Sciences and Engineering Research Council, CRL Keihanna Info-Communications Research Laboratories, and Telecommunications Advancement Organization of Japan. We wish to thank Nina Rytwinski, Laurel Fais, and Pam Thompson for helpful comments on earlier drafts of this manuscript, Marcia Riley for her work on the animation, and Mayu Nishimura for help testing subjects.

REFERENCES

Badin, P., Bailly, G., Reveret, L., Baciu, M., Segebarth, C., & Savariaux, C. (2002). Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics*, *30*, 533–553.

Bernstein, L.E., Demorest, M.E., & Tucker, P.E. (2000). Speech perception without hearing. *Perception & Psychophysics*, *62*, 233–252.

Bernstein, L.E., Eberhardt, S.P., & Demorest, M.E. (1998). Single-channel vibrotactile supplements to visual perception of intonation and stress. *Journal of the Acoustical Society of America*, *85*, 397–405.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*, 141–201.

Cutler, A., & Otake, T. (1994). Mora or phoneme? Further evidence for language-specific listening. *Journal of Memory and Language*, *33*, 824–844.

Cutler, A., & Otake, T. (1999). Pitch accent in spoken-word recognition in Japanese. *Journal of the Acoustical Society of America*, *105*, 1877–1888.

Fisher, C.G. (1969). The visibility of terminal pitch contour. *Journal of Speech and Hearing Research*, *12*, 379–382.

Graf, H.P., Cosatto, E., Strom, V., & Huang, F.J. (2002, May). *Visual prosody: Facial movements accompanying speech*. Paper presented at the 5th International Conference on Automatic Face and Gesture Recognition, Washington, DC.

Granström, B., House, D., & Lundeberg, M. (1999, August). *Prosodic cues in multimodal speech perception*. Paper presented at the XIVth International Congress of Phonetic Sciences, San Francisco.

Hadar, U., Steiner, T.J., Grant, E.C., & Rose, F.C. (1983). Head movement correlates of juncture and stress at sentence level. *Language and Speech*, *26*, 117–129.

Hadar, U., Steiner, T.J., Grant, E.C., & Rose, F.C. (1984). The timing of shifts in head posture during conversation. *Human Movement Science*, *3*, 237–245.

Hadar, U., Steiner, T.J., & Rose, F.C. (1984). Involvement of head movement in speech production and its implications for language pathology. *Advances in Neurology*, *42*, 247–261.

Hill, H., & Johnston, A. (2001). Categorising sex and identity from the biological motion of faces. *Current Biology*, *11*, 880–885.

Horn, B.K.P. (1987). Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America*, *4*, 629–642.

House, D. (2002, September). *Intonational and visual cues in the perception of interrogative mode in Swedish*. Paper presented at the 7th International Conference on Spoken Language Processing, Denver, CO.

House, D., Beskow, J., & Granström, B. (2001, September). *Timing and interaction of visual cues for prominence in audiovisual speech perception*. Paper presented at Eurospeech 2001, Aalborg, Denmark.

Jiang, J., Alwan, A., Keating, P.A., Auer, E.T., & Bernstein, L.E. (2002). On the relationship between face movements, tongue movements and speech acoustics. *EURASIP Journal on Applied Signal Processing*, *11*, 1174–1188.

Kuratate, T., Yehia, H., & Vatikiotis-Bateson, E. (1998). Kinematics-based synthesis of realistic talking faces. In D. Burnham, J. Robert-Ribes, & E. Vatikiotis-Bateson (Eds.), *International Conference on Auditory-Visual Speech Processing (AVSP'98)* (pp. 185–190). Terrigal-Sydney, Australia: Causal Productions.

Massaro, D.W., Tsuzaki, M., Cohen, M.M., Gesi, A., & Heredia, R. (1993). Bimodal speech perception: An examination across languages. *Journal of Phonetics*, *21*, 445–478.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Nicholson, K.G., Baum, S., Cuddy, L., & Munhall, K.G. (2002). A case of impaired auditory and visual speech prosody perception after right hemisphere damage. *Neurocase*, *8*, 314–322.

Risberg, A., & Lubker, J. (1978). Prosody and speechreading. *Speech Transmission Laboratory Quarterly Progress Report and Status Report*, *4*, 1–16.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, *90*, 1797–1805.

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, *21*, 427–444.

Soto-Faraco, S., Sebastián-Gallés, N., & Cutler, A. (2001). Segmental and suprasegmental mismatch in lexical access. *Journal of Memory and Language*, *45*, 412–432.

Srinivasan, R., & Massaro, D. (2002). *Synthesis and perception of prosody.* Unpublished manuscript, University of California, Santa Cruz.

Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212–215.

Thompson, D.M. (1934). On the detection of emphasis in spoken sentences by means of visual, tactual, and visual-tactual cues. *Journal of General Psychology*, *11*, 160–172.

Yehia, H.C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, *30*, 555–568.

Yehia, H.C., Rubin, P.E., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*, 23–44.

(Received 10/14/02; Revision accepted 3/21/03)