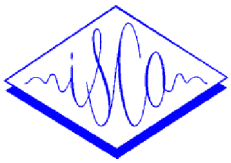

Auditory-Visual Speech Processing
2005 (AVSP'05)
British Columbia, Canada
July 24-27, 2005

COGNITIVE PROCESSING OF AUDIOVISUAL CUES TO PROMINENCE

Marc Swerts and Emiel Krahmer

Tilburg University, The Netherlands

ABSTRACT

There is growing evidence that speakers use both *auditory* markers (such as pitch accents or increased syllable durations) and *visual* markers (such as head nods and eyebrow movements) to indicate which words in an utterance are more important than others. In general, it appears that the auditory markers have stronger cue value from an observer's point of view than the visual ones, but it has also become clear that the latter have a strong impact as well (e.g., [1, 2]). Experiments with incongruent stimuli (i.e., manipulated utterances in which auditory and visual markers of prominence are associated with different words) reveal that these lead to more confusion among perceivers when they are asked which word in an utterance is the most prominent one [3]. Moreover such incongruencies are disliked by observers, presumably because they are unnatural [4]. However, we still need a good deal of knowledge on how people process combinations of auditory and visual markers to prominence.

The current paper addresses two related questions regarding prominence perception:

- How important are facial features compared to auditory ones?
- Which facial areas are most important to signal prominent words?

To address these questions we performed one production experiment and two perception experiments. In the production experiment, eight native speakers of Dutch were instructed to utter the Dutch sentence "Maarten gaat maandag naar Mali" (i.e., *on monday Maarten goes to Mali*) in a number of different conditions, each time with emphasis on one of the words (Maarten, maandag or Mali). All utterances were recorded with a digital camera (front view of the head). A selection from these recordings was used for the two perception experiments.

The first perception experiment addressed the relation between auditory and visual cues by means of a reaction-time experiment. For this experiment, recordings collected in the production experiment were systematically manipulated (mixing the speech from one utterance with the facial expressions of another utterance) in such a way that auditory and visual accents were either congruent (occurring on the same word) or incongruent (in that the auditory and the visual accent were positioned on different words). Subjects were instructed to indicate as fast as possible which word they perceived as the most prominent one. Results show that incongruent stimuli lead to slower reaction times than congruent stimuli, when looking only at instances in which participants perceived the word with the auditory accent as the most prominent one. This shows that subjects are sensitive to visual information to prominence, even in cases where they do not use this information in their actual choice.

The second experiment investigates which area of a speaker's face contains the strongest cues to prominence. We presented subjects (different from those participating in the first experiment) with film fragments of speakers collected in the production experiment. Sound and video of the utterances were artificially mixed and manipulated, this time to create stimuli with monotonous pitch, but with a visual accent on either the first, second or third noun phrase. In addition, besides stimuli showing the entire face, we also displayed stimuli in which parts of the face were made invisible, so that participants could either see only the upper or lower half, or the right or left part of the face. To compensate for potential ceiling effects, subjects were positioned at a distance of either 50cm, 250cm or 380cm from the screen which displayed the film fragments. The task of the subjects was to indicate for each stimulus which word they perceived as the most prominent one. Results show that, while prominence detection becomes more difficult at longer distances, the upper facial area has stronger cue value for prominence detection than the bottom part, and that

the left part of the face is more important than the right part. We are currently exploring to what extent the results in the horizontal axis are due to speaker (speakers might be more expressive with the left side of their face, as suggested by, e.g., Bolinger [5]) or observer effects (observers may focus more on the left side of the face, e.g., [6]), by redoing the original experiment with all stimuli mirror-reversed.

REFERENCES

[1]. Keating, P., Baroni, M., Mattys, S., Scarborough, R., Alwan, A., Auer, E., & Bernstein, L. (2003). Optical phonetics and visual perception of lexical and phrasal stress in English., in: *Proceedings of the International Conference of Phonetic Sciences (ICPhS)*, (pp. 2071--2074). Barcelona, Spain

[2]. Dohen M, Lœvenbruck H, Cathiard M.-A. & Schwartz J.-L. (2004). Visual perception of contrastive focus in reiterant French speech. *Speech Communication* **44**, 155-172.

[3]. Swerts, M. & Krahmer, E. (2004), Congruent and incongruent audiovisual cues to prominence, in *Proceedings of Speech Prosody 2004* (pp. 271--274). Nara, Japan.

[4]. Krahmer, E. & Swerts, M. (2004), More about brows, in: Zs. Ruttkay and C. Pelachaud (Eds.), *From brows to trust: Evaluating Embodied Conversational Agents* (pp. 191--216). Dordrecht: Kluwer Academic Press.

[5]. Bolinger, D. (1985), *Intonation and its parts*, London: Edward Arnolds.

[6]. Thompson, L., J. Malmberg, N. Goodell, R. Boring (2004). The distribution of attention across a talker's face, *Discourse Processes* **38**(1):145-168.