Language and Speech

http://las.sagepub.com

Audiovisual Prosody Introduction to the Special Issue

Emiel Krahmer and Marc Swerts

Language and Speech 2009; 52; 129
DOI: 10.1177/0023830909103164

Published by:
SAGE
http://www.sagepublications.com

Audiovisual Prosody—Introduction to the Special Issue

Emiel Krahmer, Marc Swerts

Tilburg University

It is a somewhat surprising fact that the vast majority of prosody research in the past has focused solely on the auditory modality. This is surprising since such a unimodal perspective is arguably not fully representative for the most archetypical communicative situation, namely a face-to-face setting in which both speaker and addressee see and hear each other, and continuously pay attention to both auditory and visual cues (e.g., Clark & Krych, 2004).

Of course, it has long been recognized that visual information (in particular lip movements) is important for speech perception, as is well known from, for example, the seminal McGurk effect (McGurk & MacDonald, 1976). Gradually researchers have started to realize that visual speech not only involves the lips, but also the rest of the face. Munhall et al. (2004), for instance, showed that auditory speech perception is improved when head movements are taken into account. And moreover, people realized that a speaker's face (and perhaps even the rest of the body) not only contributes to speech understanding, but also helps for traditional prosodic functions such as phrasing and emphasis.

A parallel trend was that researchers working on audiovisual speech synthesis became interested in the visual support of speech as well. Obviously, such audiovisual synthesis requires adequate lip movement animations (e.g., Benoît and Le Goffe, 1998, Beskow, 1995; Massaro, 1998), but a talking head only moving its tongue and lips is rather unnatural. Hence also from this research perspective, researchers have started exploring ways in which audiovisual speech can be supported with appropriate facial movements to indicate, for instance, what the information structure of the current utterance is and which words in it are the most prominent ones (e.g., Cassell, Sullivan, Prevost, & Churchill, 2000; Granström, House, & Lundeberg, 1999, Granström, House & Swerts 2002; Pelachaud, Badler, & Steedman, 1996).

Address for correspondence. Tilburg centre for Creative Computing (TiCC), Department of Communication and Information Sciences (DCI), Tilburg University, The Netherlands; e-mail: <e.j.krahmer, m.g.j.swerts}@uvt.nl>

Language and Speech 0023-8309; Vol 52(2/3): 129-133; 103164; DOI:10.1177/0023830909103164
http://las.sagepub.com

As a result of these two converging trends, the visual modality has recently started to receive more attention in the study of prosody. Due to better and cheaper equipment for recording and storing data, and due to increased computing power for automatic analyses, this is now much easier than even a few years ago. The first results of this kind of research strongly suggest that the visual component indeed has a clear added value for various aspects of communication that in the past were typically associated with verbal prosody. Various researchers have looked at visual correlates of prominence and focus (e.g., Cavé, Guaïtella, Bertrand, Santi, Harlay, & Espessev, 1996; Erickson, Fujumura, & Pardo, 1998; Hadar, Steiner, Grant, & Rose, 1983; Krahmer & Swerts, 2004, 2007; Swerts & Krahmer 2008), showing, for instance, that visual cues such as eyebrow flashes, head nods, and beat gestures boost the perceived prominence of the words they occur with, and downscale that of the surrounding words. In a similar vein, audiovisual cues for such traditional prosodic functions as phrasing (e.g., Barkhuysen, Krahmer, & Swerts, 2008), face-to-face grounding (e.g., Nakano, Reinstein, Stock, & Cassell, 2003), and question intonation (e.g., Srinivasan & Massaro, 2003) have been explored, as have the audiovisual expressions of affective functions such as signaling basic emotions (e.g., Barkhuysen, Krahmer, & Swerts, 2009, among many others and de Gelder et al., 1999) and social ones like uncertainty (Krahmer & Swerts, 2005; Swerts & Krahmer, 2005) and frustration (Barkhuysen, Krahmer & Swerts, 2005).

It is thus fair to say that audiovisual information has been shown to be important for a wide range of communicative functions, as they may influence both speech intelligibility and signal higher-level pragmatic issues (like emotion and attitude). Audiovisual prosody thus serves a clear purpose in human–human interactions, and there is growing evidence that it may make human–machine interactions more effective as well. However, to be fair, work on audiovisual prosody up to now has only addressed a limited number of topics and many avenues of further research have not been explored. A special issue on this topic thus seems to be both timely and important. The papers that have been selected cover multiple perspectives on audiovisual prosody, ranging from (experimental and uncontrolled) data of real human interactions to more application-oriented approaches, and covering both audiovisual expressions of human and of artificial speakers.

We start with a quartet of papers addressing audiovisual prosody in its strictest interpretation. Scarborough, Keating, Mattys, Cho, Alwan, and Auer, Jr. ask which visual cues contribute to the perception of lexical and phrasal stress. They collected audiovisual data from various native speakers of American English, and in the visual domain a large variety of facial movements were measured, which were mostly larger and faster when associated with stressed words. It is suggested that chin measures (associated with mouth openings) are the strongest predictor of perceived stress. Dohen and Lœvenbruck study prosodic contrastive focus and its audiovisual realization in French. In many cases, it is straightforward to determine prosodic focus in auditory speech, which makes it difficult to adequately judge the added value of visual information. To deal with such ceiling effects, Dohen and Lœvenbruck propose to use whispered speech and show that this is indeed an effective paradigm to study the audiovisual expression of prosodic focus. Guaïtella, Santi, Lagrue, and Cavé address the function of rapid eyebrow movements in spoken French. In earlier work (e.g., the aforementioned Cavé et al., 1996) the authors have zoomed in on the link between eyebrow movements and changes in fundamental frequency. In this study they show that speakers' eyebrow flashes are associated *both* with turn taking and with changes in the fundamental frequency. Since many eyebrow movements occur before or early in a turn, the authors argue that eyebrow movements more often act as a turn-getting device than as a turn-holding cue or a visual accentuation indicator. Rilliard, Shochi, Martin, Erickson, and Aubergé look at the audiovisual production and perception of social affect. They do so in a comparative study of both Japanese and French, and their work reveals interesting similarities—and differences—in use and understanding of the different modalities. In general (and in line with earlier findings on different functions of audiovisual prosody), their work reveals a perceptual advantage of multimodal presentations over unimodal ones, but also substantial individual variation in expressivity and strategy.

This first quartet is followed by a trio of papers looking at visual prosody in sign languages, which offer an intriguing counterpart to the other papers in this issue. Naturally, there is no "audio" component in sign languages (at least not as it is understood in non-signed, spoken languages). In a way, the "articulatory effort" in sign languages has shifted to the hands, but the interesting question is whether the visual prosody (eyebrow movements, eye blinks, etc.) works in similar ways across signed and non-signed languages. Wilbur addresses this question for American Sign Language (ASL), with a general focus on the effects of changes in signing rate on signs, pauses and non-manual (i.e., facial) markers. Dachkovsky and Sandler explore similar issues in Israeli Sign Language (ISL). They argue that different visual signs in ISL indeed have specific prosodic functions, such that the combination of these visual signs gives rise to a subtle yet meaningful layer on top of the signing, much like intonation relates to words and sentences in spoken language. De Vos, van de Kooij, and Crasborn, finally, study the interaction between prosodic and affective functions of eyebrow movements. They do so in Sign Language of the Netherlands (NGT), using a paradigm in which signers are asked to combine different prosodic functions (signaling content or polar questions) with different affective states. A detailed analysis in terms of FACS' eyebrow-related Action Units reveals various interesting patterns in the collected data.

We have already noted above that the study of audiovisual prosody is closely related to that of audiovisual speech, and hence we decided to include one paper addressing a fundamental issue in audiovisual speech. In a short paper, Vroomen and Baart report on their findings concerning the duration of what is known as recalibration effects. When listeners hear an ambiguous speech sound, they may modify their phonetic categories in a flexible manner based on lipread information. Vroomen and Baart tested the stability and duration of such lipread-based recalibrations, arguing that they are much more fragile than has previously been assumed, which sheds an interesting light on the exact relation between auditory and visual speech.

The last two papers in this issue address tools and methods that are especially relevant for the study of audivisual prosody. Edlund and Beskow describe an experimental platform, which they dubbed "MushyPeek", in which pairs of participants communicate via a VoIP telephone connection and simultaneously see an avatar representing their communication partner. In this way it is possible to manipulate the audiovisual behavior of conversation partners in a manner that is similar to the face-to-face setting. An evaluated proof of concept implementation shows the feasibility of this approach. Finally, Theobald, Matthews, Mangini, Spies, Brick, Cohn and Boker describe a comparable set-up, but instead of avatars they use the actual visual appearance of the conversation partners. They describe a set of techniques that enable researchers to manipulate these images in real time and at video frame rate. Again, a proof of concept implementation of the approach is evaluated, with promising results.

Even from this short overview, the reader may already have noticed that a lot of ground is covered in this special issue. Data from a large variety of languages is discussed (including Dutch, French, English, and Japanese), and besides spoken languages also various signed languages (ISL, ASL, NGT) are studied. Orthogonal to this, a wide range of audiovisual prosodic cues is discussed and a correspondingly wide range of functions (ranging from lipread calibration up to turn taking, prominence signaling and affect). In this way, the articles selected for this special issue give an excellent overview of current activities in the study of audiovisual prosody, and we hope they may serve as a source of inspiration for future work into the interplay between auditory and visual correlates of prosody.

Before turning to the selected papers themselves, we would especially like to thank the many expert reviewers for their constructive criticisms on the submitted papers. We could not have done without you.

References

BARKHUYSEN, P., KRAHMER, E., & SWERTS, M. (2005). Problem detection in human–machine interactions based on facial expressions of users. *Speech Communication*, **45**(3), 343–359.

BARKHUYSEN P., KRAHMER, E., & SWERTS, M. (2008). The interplay between the auditory and visual modality for end-of-utterance detection. *Journal of the Acoustical Society of America*, **123**(1), 354–365.

BARKHUYSEN P., KRAHMER, E., & SWERTS, M. (2010). Cross-modal and incremental perception of audiovisual cues to emotional speech. *Language and Speech*, in press.

BENOÎT, C., & LE GOFF, B. (1998). Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP. *Speech Communication*, **26**, 117–129.

BESKOW, J. (1995). Rule-based visual speech synthesis. In *Proceedings of the 4th European conference on speech communication and technology (EUROSPEECH 95)* Madrid (pp.299–302).

CASSELL, J., SULLIVAN, J., PREVOST, S., & CHURCHILL, E. (2000). *Embodied conversational agents*. Cambridge, MA: MIT Press.

CAVÉ, C., GUAÏTELLA, I., BERTRAND, R., SANTI, S., HARLAY, F., & ESPESSER, R. (1996). About the relationship between eyebrow movements and F0 variations. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)* Philadelphia (pp.2175–2179).

CLARK, H. & KRYCH, M. (2004), Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, **50**, 62–81.

De GELDER, B., BÖCKER, K., TUOMAINEN, J., HENSEN, M., & VROOMEN, J. (1999). The combined perception of emotion from face and voice: Early interaction revealed by human electric brain responses. *Neuroscience Letters*, **260**, 133–136.

ERICKSON, D., FUJIMURA, O., & PARDO, B. (1998). Articulatory correlates of prosodic control: Emotion and emphasis. *Language and Speech*, **41**(3–4), 399–417.

GRANSTRÖM, B., HOUSE, D., & LUNDEBERG, M. (1999). Prosodic cues to multimodal speech perception. In *Proceedings 14th International Conference of the Phonetic Sciences (ICPhS)*, San Francisco.

GRANSTRÖM, B., HOUSE, D., & SWERTS, M. (2002). Multimodal feedback cues in human–machine interactions. In *Proceedings of Speech Prosody 2002* Aix en Provence, France (pp.347–350).

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.

KRAHMER, E. & SWERTS, M. (2004). More about brows: a cross-linguistic analysis-by-synthesis study. In C. Pelachaud & Z. S. Ruttkay (Eds.), *From brows to trust: Evaluating embodied conversational agents*, Kluwer Academic Publishers.

KRAHMER, E. & SWERTS, M. (2005). How children and adults produce and perceive uncertainty in audiovisual speech. *Language and Speech*, **48**(1), 29–54.

KRAHMER, E. & SWERTS, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, **57**(3), 396–414.

MASSARO, D. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*, Cambridge, MA: The MIT Press.

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

MUNHALL, K., JONES, J., CALLAN, D., KURATATE, T., & VATIKIOTIS-BATESON, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, **15**, 133–137.

NAKANO, Y. I., REINSTEIN, G., STOCKY, T., & CASSELL, J. (2003). Towards a model of face-to-face grounding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)* Sapporo, Japan.

PELACHAUD, C., BADLER, N., & STEEDMAN, M. (1996). Generating facial expressions for speech. *Cognitive Science*, **20**, 1–46.

SRINIVASAN, R. & MASSARO, D. (2003). Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech*, **46**, 1–22.

SWERTS, M. & KRAHMER, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, **53**(1), 81–94.

SWERTS, M. & KRAHMER, E. (2008). Facial expressions and prosodic prominence: Comparing modalities and facial areas. *Journal of Phonetics*, **36**(2), 219–238.