Visual prosody of newsreaders

Swerts, M.G.J.; Krahmer, E.J.

Published in:
Journal of Phonetics

Publication date:
2010

Document Version
Publisher's PDF, also known as Version of record

Journal of Phonetics 38 (2010) 197–206

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.de/phonetics

Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions

Marc Swerts *, Emiel Krahmer

Tilburg University, Faculty of Humanities, Tilburg Centre for Creative Computing (TiCC), Department of Communication and Information Science, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands

ARTICLE INFO

Article history:
Received 26 January 2009
Received in revised form
7 October 2009
Accepted 9 October 2009

ABSTRACT

This article investigates whether newsreaders exploit their expressive style for signalling different types of communicatively relevant information. The study reported upon here investigates whether these speakers use their facial expressions to "package" the content of their messages so that they reflect the relative importance, the emotional connotation and the intended audience of the news items. To this end, Dutch newsreaders (addressing audiences consisting of either adults or children) were analysed in terms of their facial expressions. The first study explores whether eyebrow movements and head nods of newsreaders (only adult news) are correlated with the relative prominence of words in their messages. Consistent with findings in the literature for other styles of speech, it was found that pitch-accented words in the newsreader data tended to be marked by variations in eyebrow movements and head nods. The second study aimed to find out whether newsreaders adapt their facial expressions to the seriousness of the topic they are talking about, and, if so, whether this adaptation differs for newsreaders addressing adults or children. Analyses reveal that both the topic and the intended audience had an effect of the newsreaders' expressions, especially as the positive topics and the child-directed news items tended to be more expressive. The relevance of the findings of studies 1 and 2 for speech production and recognition is discussed.

1. Introduction

1.1. Optical phonetics

An addressee derives different types of functionally relevant information from the visual appearance of a speaker. In particular, dynamic variation in the speaker's face has been shown to have strong effects on how an incoming message is interpreted. At the segmental-phonological level, there is a great deal of evidence that facial cues affect speech intelligibility. Both people who have a hearing problem and those with normal hearing can understand a speaker better if they can see that speaker's lip movements, and, interestingly, also when they see rhythmic movements in the upper part of his or her face (Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). Such visual information is especially useful in noisy conditions (Sumby & Pollack, 1954), or when a speaker is addressing someone who is relatively distant in space (Jordan & Sergeant, 2000). In addition, rather than influencing speech intelligibility, a speaker's face may also signal extra information that is not necessarily expressed through the words or syntactic structure of a sentence. The face may "qualify" a spoken message in that it may for instance reveal whether information is important to the discourse or not, or what a person's attitude is regarding the message (e.g., he or she may be ironic about its content), or whether the speaker is emotionally affected by the content. In this respect, the facial expressions, together with other forms of body language, constitute a form of visual prosody (Graf, Cosatto, Strom, & Huang, 2002). That such expressions are useful for human interactions is very clear from the fact that text chatters abundantly use emoticons like :-) to signal elements of information that may not be clear from the words and syntax alone.

The increased awareness about the relevance of the visual modality, viz. the face, has led to a new discipline which is sometimes referred to as "optical phonetics" (Scarborough et al., 2009). An important problem for this field concerns the relationship between auditory and visual information in human communication. There is growing evidence showing that the two modalities are very much related, in the sense that elements of information that have been shown to be signalled by auditory cues may also be transmitted through facial expressions. However, while the validity of facial cues for speech intelligibility at the segmental level has repeatedly been shown, there is still a lot that needs to be learned about the way facial cues are exploited for qualifying the informational content, and how such cues relate to

* Corresponding author. Tel.: +31 13 4662922; fax: +31 13 4663110.
E-mail address: m.g.j.swerts@uvt.nl (M. Swerts).

0095-4470/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2009.10.002

auditory prosodic signals. Regarding the latter, there are basically two sets of questions, depending on whether one takes a production- or perception-oriented perspective, that await further research. First, to what extent do speakers exploit facial expressions for marking elements of information that are also encoded in auditory information? And second, how sensitive are addressees to variation in the face regarding aspects that may also be signalled by auditory variables?

Unfortunately, so far, there has been some concern about the empirical basis of the studies that address such issues. Approaches that base their analyses on spontaneous speech data have often been criticized for using anecdotal evidence. To counter such problems, there have been a number of laboratory studies into the functional use of facial expressions, either by eliciting such expressions from real speakers in specifically designed linguistic or social contexts (Dohen, 2005; Dohen & Lœvenbruck, 2009; Dohen, Lœvenbruck, Cathiard, & Schwartz, 2004; Krahmer & Swerts, 2007; Scarborough et al., 2009; Swerts & Krahmer, 2008), or by using an analysis-by-synthesis procedure whereby the effect of specific variations in the face on observers are tested through controlled manipulations of synthetic talking heads (Cassell, Vihjálmsson, & Bickmore, 2001; Granström, House, & Lundeberg, 1999; Krahmer, Ruttkay, Swerts, & Wesselink, 2002; Krahmer & Swerts, 2004; Pelachaud, Badler, & Steedman, 1996). Such studies showed that facial expressions are indeed functionally relevant as they are exploited to support a broad range of communicative functions (e.g., to highlight important information, or to reveal attitudinal or emotional correlates of speaker utterances). While such experimentally controlled set-ups will allow a researcher to test very specific hypotheses, there is also a danger that elicited and manipulated data lead to discovering experimental artefacts, i.e., findings that are mainly true for the laboratory contexts, but do not necessarily generalize to more natural settings (Wilting, Krahmer, & Swerts, 2006). Moreover, a large proportion of the studies tend to be speaker-oriented, so that it remains to be seen whether the facial patterns produced by speakers are also perceptually relevant for addressees. Sometimes such a perceptual validation is not self-evident, as the production studies make use of recordings with speakers who have special markers on their faces which facilitate (automatic) analyses of the expressions, but which render the faces less suitable for perception tests. In other words, a critical issue, which has hampered many studies in this area of research, is to establish a good empirical basis for claims regarding the functional use of a speaker's visual cues, while at the same time it remains important that the data allow the controlled testing of specific hypotheses about facial cues.

The current study analyses recordings of professional newsreaders. Such recordings represent natural data that are still sufficiently constrained to be able to explore specific functions of their expressive style. On the one hand, newsreaders are often viewed as role models as they represent the "best" speakers within a linguistic community; accordingly, newsreader data have often served as a basis for models of automatic speech synthesis and recognition. On the other hand, the newsreader data tend to be topically constrained and the conditions under which they are produced remain fairly constant. The purpose of the research presented here is to provide an answer to questions about the relative importance of visual information from the face for "information packaging".

1.2. Information packaging

Chafe (1976) originally introduced the notion of "information packaging" to cover a range of phenomena that have to do primarily with how the message is sent and only secondarily with the message itself, just as the packaging of toothpaste can affect sales in partial independence of the quality of the toothpaste inside. Prosodic features are an important aspect of information packaging, as they have been shown to be exploited for organizing the information. Indeed, past research on auditory prosody has repeatedly shown that speakers use auditory features, like intonation, temporal variation, and voice quality, to mark aspects of information which go beyond the mere propositional content of their messages. Here, it will be investigated to what extent facial expressions may support functions that have previously been shown to be served by auditory prosodic patterns. The current study will explore whether facial expressions may reveal (1) which elements in the discourse are relatively more important, (2) whether a news item relates to a positive or negative event and (3) whether newsreaders adapt their expressions to specific audiences.

First, many researchers of spoken discourse argue that speakers use specific linguistic devices to distinguish elements of information that are central to the discourse from elements that belong to the background. In particular, for many languages it has been claimed that speakers and listeners make use of pitch accents and other prosodic variables to highlight important pieces of information in their utterances. In addition, there is growing evidence that speakers also use particular facial expressions and other bodily gestures for the same purpose. In particular, so-called beat gestures, i.e., quick movements of the eyebrows, the head or the hand, that a speaker produces while uttering a sentence, may mark which words are the more prominent ones in an utterance. So far, production experiments show that speakers may indeed accompany their pitch-accented words by co-varying facial information and hand gestures (Cavé et al., 1996; Flecha-Garcia, 2006a, 2006b, 2007; Graf et al., 2002; Scarborough et al., 2009; Yasinnik, Renwick, & Shattuck-Hufnagel, 2004). Perception studies with carefully controlled (Dohen & Lœvenbruck, 2009; Krahmer & Swerts, 2007), manipulated (Swerts & Krahmer, 2008) or synthesized stimuli (Krahmer et al., 2002; Krahmer & Swerts, 2004) report evidence that observers are sensitive to such visual cues. While the auditory cues to prominence appear to be more dominant than the visual cues, viewers nevertheless find stimuli more natural and react quicker when the auditory and visual cues to accents are congruent (i.e., occur on the same word) rather than incongruent. In addition, visual signals boost the perceived prominence of accented words, and diminish the perceived prominence of neighbouring words.

The results of the studies presented above are almost exclusively based on highly constrained, experimental data, where speakers are sometimes even explicitly instructed to accent specific words. In contrast, Yasinnik et al. (2004) and Flecha-Garcia (2006a, 2006b, 2007) present analyses of recordings of speakers who are speaking more freely. The former looked at speech-accompanying gestures in recordings from three academic lecturers, finding that discrete gestures of head and hands may be timed with respect to pitch accented syllables (or possibly the prominence-related constituents they define). The latter examined eyebrow raises in face-to-face interactions (Scottish English Maptask data). In particular, Flecha–Garcia studied eyebrow raises in relation to discourse structure and utterance function (e.g., questions versus instructions), and investigated whether they correlate with pitch accents. One of her findings which is particularly relevant for our current study is that in her data pitch accents occur in coordination with eyebrow raises. She found statistical evidence that brow raises began significantly closer to pitch accents surrounding them (especially following accents) compared to a random distribution. But, as Flecha–Garcia concludes herself, it still remains to be seen whether such findings

generalize to other speakers and other types of discourse as well, such as the monologues we want to explore in the current study. One important difference between her and the newsreader data is that newsreaders use a rather formal speaking style and may therefore not be as expressive as speakers in the spontaneous interactions of Flecha–Garcia. This is potentially relevant as it has been shown that facial movements on accents may differ for utterances in different expressive modes (Beskow, Granström, & House, 2006). In addition, in contrast with the studies cited above, our data will show the extent to which eyebrow movements co-occur with another type of gesture, head nods. And finally, our research focuses on a different language, Dutch, which is important since the implementation of prominence may be different in Dutch than in other languages (Ladd, 1996; Krahmer & Swerts, 2004).

Second, the informational status of news items may also vary along another dimension. Utterance fragments may not only differ regarding their informational importance in the ongoing discourse, but may also relate to topics which have different emotional connotations. For instance, a speaker may refer to a very sad or rather joyful event. Already in the 1st century AD, Quintilianus remarked in his *De institutio oratoria* that a speaker's expressive style should be congruent with the content of what he or she is saying, since "if we look cheerful when our words are sad, or shake our heads when making a positive assertion, our words will not only lack weight, but will fail to carry conviction".[1] In practice, however, we still lack sufficient knowledge about the extent to which speakers indeed adapt their expressive style to the emotional content of their messages. Of course, starting with Darwin, it has repeatedly been shown that facial expressions can signal a speaker's emotion, like anger, fear, or happiness. In addition, these emotions are also reflected in specific prosodic features, such as intonation, speech rate, loudness or voice quality. But again, results from such analyses have been based on data that sometimes have a questionable ecological validity. Regarding facial analyses, the experiments often make use of still images, like photographs or drawings, that may not be representative of the way people perceive "moving" faces in natural interactions. Similarly, for methodological and ethical reasons, both visual and auditory analyses tend to make use of recordings of actors who are invited to produce different kinds of emotions in a studio. Unfortunately, there is evidence that posed expressions are markedly different from expressions during spontaneous interactions, as the acted emotions are more extreme and more stereotypical than the natural ones (Wilting et al., 2006). Accordingly, there is a growing interest in the development of methods to spontaneously elicit different kinds of emotions, but so far these have not been widely applied within the scientific community.

And finally, there is a third way in which the expressive style of speakers may be exploited for communicative purposes. Possible visual cues to information status and emotional content are triggered by characteristics of the message itself. In addition, it could be that the newsreaders' style is dependent on some contextual information, in particular the kinds of viewers they are addressing. There is a whole body of evidence to suggest that speakers adapt their general communicative style to specific characteristics of their addressees. Variation due to such forms of audience design (Clark, 1992) may occur on different linguistic levels: e.g., when addressing a stranger, we tend to use a different set of words and syntactic structures than in an interaction with a family member. Adaptation in terms of auditory features like speech rate and pitch range appears to be especially important in interactions with addressees who have less developed communicative skills, like children, non-native speakers and pets (Burnham, Kitamura, & Vollmer-Conna, 2002) or in interactions with a poorly performing spoken dialogue system (Hirschberg, Litman, & Swerts, 2004; Oviatt, Bernard, & Levow, 1999). It is generally found that adults, in their interactions with these kinds of conversation partners, use a comparatively more expressive style (involving hyperarticulation). While most of the evidence about audience design is based on auditory information, little is known about the extent to which facial expressions are tailored to particular addressees. There is a study by Barkhuysen, Krahmer, and Swerts (2005) which showed that speakers tend to produce more marked facial expressions in their interaction with a "bad" spoken dialogue system, especially when they notice that the interaction is not running smoothly. Along the same lines, it is interesting to explore whether a speaker's expressive style differs depending on whether he or she is talking to an adult or a child.

1.3. Research questions

The overview of studies above showed that speakers can vary their expressive style for a range of communicative purposes. In particular, past research has shown that speakers exploit auditory variables to package the information according to at least three factors: (1) the relative importance of discourse unit, (2) the emotional content of the topics that are addressed and (3) the intended audience. However, based on our discussion above of previous research, there are reasons to assume that these factors may also have an impact on the facial expressions of speakers. The current investigation looks in to the extent to which three potentially relevant factors, i.e., the distribution of pitch accents, emotional content and intended audience, have an effect on the visual appearance of newsreaders. The article reports about two studies. The first study explores different newsreaders from Dutch public TV, and investigates whether words that are prosodically highlighted in their utterances are also marked by rapid eyebrow movements and head nods. To this end, the study correlates auditory and visual mark-ups of recorded news bulletins. The second study studies whether the emotional content of a message, i.e., whether it represents "good" or "bad" news, is visually signalled by newsreaders to their viewers. In doing so, it compares readers who present these two kinds of news items to two kinds of audience, namely adults versus children, to see whether this difference is reflected in their general expressive style. The questions of the second study are addressed by two perception experiments in which participants are asked to judge news items (from adult and child news) in terms of their expressiveness and emotional content.

2. Study 1

2.1. Goal

The first study explores whether facial expressions (both eyebrow movements and head nods) of newsreaders are correlated with the relative prominence of words in their messages.

2.2. Method

2.2.1. Materials

The creation of the database with audiovisual recordings was organized as follows. In January 2007, four Dutch newsreaders (two male speakers: Philip Freriks, Rick van de Westelaken; two female speakers: Jeanet Schuurman, Sacha de Boer) were recorded

[1] Translation adapted from http://www.archive.org.

Fig. 1. Representative still of one of the newsreaders (Sacha de Boer) we analysed, while she was reading the adult-directed news.

while they were reading the 8 o'clock news on Dutch public TV. Fig. 1 shows a representative still of a recording of one of the analysed newsreaders (Sacha de Boer). Of all those recordings, 60 fragments were randomly selected, with an equal number of fragments of the four speakers. If it turned out that a fragment happened to contain a speech dysfluency or other anomalous pronunciation, that fragment was replaced with another one. The fragments consisted of one full sentence, the shortest lasting 4 s, the longest 12 s. All the utterances were transcribed orthographically so that they could be used in the labelling sessions.

2.2.2. Mark-up

The selected materials were then annotated in terms of prominence structures, and in terms of visual signals from the newsreaders (eyebrow movements and head nods).

Auditory mark-up: The first was done by means of a rating experiment with 35 participants (14 male, 21 female) between 18 and 54 years old (average: 24.2). In order to make sure that the visual analyses were independent from the auditory ones, we first extracted the wave form (the auditory information) from the news recordings. Via an internet application, the waveforms per sentence were randomly presented to the participants, each participant receiving a different random order to compensate for possible learning effects. The task given to the participants was to determine which word or which words in each sentence were prosodically highlighted, which was informally defined as words which are produced with clear auditory emphasis. Listeners could listen as often as they wished to the sentences, but after they had made a decision they could not return to an earlier sentence. Each listening session (including reading the instructions) lasted approximately 30 min per participant.

For subsequent statistical analyses, the prominence ratings from the 35 annotators were categorized into a smaller set, following a general procedure that has previously been used for discourse and prosodic marking (e.g., Swerts, 1997; Mo, Cole, & Lee, 2008). More specifically, based on the idea that cue strength is related to the proportion of subjects that agrees on a label, a word was categorized as having no accent if none of the participants had annotated it as having an accent, whereas a word which at least half of the participants had indicated as being prominent was labelled as a strong accent; the cases in between those two extremes were classified as having a weak accent. Following this procedure, we found that of the total of 985 words in the 60 sentences, 609 (62%) had "no accent", 309 (31%) had a "weak accent", and 67 (7%) had a "strong accent".

Visual markup: For the visual annotations, we used the Noldus™ Observer™ software, which was designed as a tool to analyse behavioural observations. The sentences were visually annotated by two independent researchers, different from the authors, who in a first round did the annotations separate from each other, where it took them about 8 min per sentence to label the data. After the first round, the two separately obtained annotations were compared, and those sentences on which there was disagreement were viewed again in order to get a consensus labelling on presence or absence of a visual cue. The visual features that were labelled were rapid eyebrow movements and head nods. It was decided to label an event as a rapid eyebrow movement if at least one of the eyebrows changed in upward or downward direction. Following this procedure, the data contained 303 eyebrow movements. It turned out that the majority of these (59%) consisted of simultaneous upward movements of both eyebrows. An event was labelled as a head nod if the head moved in right or left direction, or forward or backward. The data contained 228 head movements, of which the most frequent ones were distributed as follows: 27% were head nods to the left, 14% head nods to the right, 28% head nods forward and 26% head nods upward. The remaining analyses are based on a simple distinction of presence or absence of eyebrow movement or head nod, without further distinctions between different types.

The eyebrow movements and head nods were first annotated and compared without access to the auditory signal. After that, the visual signals were compared with the auditory channel to determine on which words the two kinds of visual gestures occurred. For most cases, the complete movement (from onset to offset) occurred within a single word. If, however, a movement happened to extend over multiple words (which happened in some rare cases), the word which contained the peak moment of the movement was labelled as being aligned with the movement.

2.3. Results

The current section will first discuss results for the eyebrow movements and head nods separately, and then consider patterns of co-occurrences. The results for eyebrow movements as a function of occurrence of different accents is given in the upper half of Table 1. This table reveals that the distribution of the eyebrow movements is statistically dependent on that of the accent types ($\chi^2 = 73.26, \mathrm{df} = 2, p < 0.001$). That is, while the majority of the "no accent" and "weak accent" cases occur without an accompanying eyebrow movement (76.8% and 62.8%, respectively), the strong accents more often co-occur with eyebrow movements (70.1%) than not. However, conversely, the mere presence of an eyebrow movement does not imply

Table 1
Distribution of eyebrow movements and head nods as a function of three levels of accent (overall results).

Movement	Level	Accent			Total
		No accent	Weak accent	Strong accent	
Eyebrow	Absent	468 (76.8%)	194 (62.8%)	20 (29.9%)	682
	Present	141 (23.2%)	115 (37.2%)	47 (70.1%)	303
	Total	609 (100%)	309 (100%)	67 (100%)	985
Head	Absent	565 (92.8%)	185 (59.9%)	7 (10.4%)	757
	Present	44 (7.2%)	124 (40.1%)	60 (89.6%)	228
	Total	609 (100%)	309 (100%)	67 (100%)	985

The table shows both absolute numbers and column percentages.

the presence of a strong accent, given that only a minority of the eyebrow movements (47 out of 303) occurs with a strong accent. In other words, while the presence of a strong accent is likely to co-occur with an eyebrow movement, the reverse is not true. A similar pattern emerges for the distribution of the head movements, as shown in the bottom half of Table 1. As with eyebrow movements, the distribution of head movements is related to that of the accent status of words ($\chi^2 = 302.97$, $df = 2, p < 0.001$). As can be seen in the table, strong accents co-occur with head movements in 89.6% of the cases. And as was the case with eyebrows, the relation between accents and head nods is not symmetrical, as only 60 out of 228 head nods appear together with strong accents.

The overall pattern in Table 1 appears to be true for each newsreader separately as well. Table 2 gives percentages for each of the three accent categories of words with eyebrow or head movements, respectively, separated for each of the four newsreaders. As is clear, the likelihood that a word co-occurs with an eyebrow or head movement is the highest for strong accents, and weakest for words without accent. This pattern is true for both visual markers, and for all four newsreaders.

The data in Tables 1 and 2 show the separate relations (i) between accents and eyebrow movements and (ii) between accents and head movements. However, it could be that the two forms of visual information themselves are not independent. Therefore, it may be revealing to consider combinations of these types of visual information as well. Table 3 shows actual numbers and percentages of words that have no visual information at all, words that have only an eyebrow movement (but no head movement), words with only a head movement (but no eyebrow movement), and words which are accompanied by both eyebrow and head movements. The table shows that strong accents are especially marked by combinations of eyebrow and head movements (67.2%), while single eyebrow or head movements appear to be far less typical for strong accents. In addition, the combined occurrences of eyebrow and head movements rarely occur on words without an accent (19 out of 138 cases).

Table 2
Percentage of words with an eyebrow movement and head nod as a function of three levels of accent.

Newsreader	Movement	Accent		
		No accent	Weak accent	Strong accent
PF	Eyebrow	25.2	26.7	57.1
	Head	5.7	37.2	100
RW	Eyebrow	31.9	65.0	83.3
	Head	8.1	51.7	83.3
JS	Eyebrow	13.7	25.8	90.0
	Head	8.2	36.0	100
SB	Eyebrow	20.8	40.5	50.0
	Head	6.9	39.2	75.0

The table lists column percentages for four newsreaders separately: Philip Freriks (PF), Rick van de Westelaken (RW), Jeanet Schuurman (JS) and Sacha de Boer (SB).

Table 3
Distribution of words without any head movement, words with only a movements of the eyebrow or only a head nod, and words with both an eyebrow movement and head nod, as a function of three levels of accent (overall results).

Movement	Accent			Total
	No accent	Weak accent	Strong accent	
No movement at all	443 (72.7%)	144 (46.6%)	5 (7.5%)	592
Only eyebrow movement	122 (20.0%)	41 (13.3%)	2 (3.0%)	165
Only head movement	25 (4.1%)	50 (16.2%)	15 (22.4%)	90
Eyebrow and head movement	19 (3.1%)	74 (23.9%)	45 (67.2%)	138
Total	609 (100%)	309 (100%)	67 (100%)	985

The table shows both absolute numbers and column percentages.

2.4. Discussion

Consistent with findings of previous studies (Cavé et al., 1996; Flecha-Garcia, 2006a, 2006b, 2007; Graf et al., 2002; Scarborough et al., 2009; Yasinnik et al., 2004), the analyses of the newsreader data in study 1 showed that newsreaders use specific forms of visual prosody, in particular joint movements of eyebrow and head, to mark certain words in their utterances as prominent. These signals appear to align with auditory correlates of prominence. The co-variation may be due to the fact that speakers have a tendency to align various acoustic and visual forms of variation. This effect to combine audiovisual information is relevant from the perspectives of both speech production and perception. Regarding production, the analyses show that speakers tend to use both auditory and visual features for marking important information in their utterances. This result is potentially useful for models of speech production that take into account the visual modality as well (Krahmer & Swerts, 2007), as our results suggest that both correlates tend to be produced in tandem by a speaker. In addition, this insight may be helpful for the development of embodied conversational agents, like synthetic talking heads. In order to make such human-like interfaces more natural, many researchers have explored procedures to introduce variation in the acoustic output and in the visual appearance of such heads. In principle, the co-occurrence of auditory and visual markers of prominence is also relevant for recognition purposes, even when this would have to be tested in follow-up studies. Previous perception studies with artificially constructed or experimentally elicited stimuli (Dohen et al., 2004; Krahmer et al., 2002; Krahmer & Swerts, 2004, 2007; Scarborough et al., 2009; Swerts & Krahmer, 2008) show that observers are indeed sensitive to visual markers of prominent information, albeit that auditory markers tend to have stronger cue value. It remains to be seen whether such results generalize to more natural data, and whether the visual markers of important information indeed aid or influence the speech decoding process. Along the same lines, it could be useful to explore whether automatically detected visual markers of prominent information, together with auditory information, could be applied to find important stretches of speech in a discourse, which in turn could be useful for speech summarization systems.

In addition, it is worthwhile to explore whether the pitch accents that are accompanied by visual information are systematically different from those that lack such a visual cue. One relevant question, for instance, is whether there is a difference in the communicative context for accents that co-occur with facial markers from those that are not supported by visual information. Looking at some examples in our recorded data suggests that it may matter whether accents signal contrastive information or not. For instance, in examples (1), (2) and (3) the capitalized word always represents an element of information which contrasts with information earlier in the sentence. Interestingly, these accents are always very prominent, and accompanied by significant visual information.

(1) "…één meer dan in het tweede kabinet Balkenende, terwijl MINDER ministers en staatssecretarissen de bedoeling was"

"…*one more than in Balkenende's second government, whereas there should have been FEWER ministers and secretaries of state*"

(2) "De politie is op zoek naar twee of DRIE verdachten"
"*The police is searching for two or THREE suspects*"

(3) "…zijn vandaag vijfentwintig doden en ruim HONDERD gewonden gevallen"
"…*today there were twenty-five dead and more than HUNDRED wounded*"

Analyses of larger sets of data are needed to find out whether the occurrence of facial information indeed depends on different types of discourse variation. Along the same lines, it would make sense to try and differentiate between types of accents, and explore whether there is a connection with presence or absence of facial information. For instance, it has been argued that the choice of using H^* or $L+H^*$ accents is related to differences in information structure, so that it would be interesting to see whether that difference is also reflected in different kinds of eyebrow gestures or other variation in the face. So far, we have restricted ourselves in the current study to a simple distinction between accented and unaccented words, but there have been claims in the literature that different accent types may serve different communicative purposes. For instance, there is an ongoing debate as to whether broad focus and narrow focus accents are distinct regarding phonological properties and accent strength (Krahmer & Swerts, 2001). Given that previous studies have shown that visual cues "boost" the perceived prominence of a pitch accent (Krahmer & Swerts, 2007; Krahmer et al., 2002), one could hypothesize that visually marked accents are especially suitable as cues to highly significant discourse information, such as prominent contrasts. Our method of determining accents may serve as a good alternative to determine accent strength, as we still lack an agreed-upon method to specify gradient differences in prominence. Alternatively, to get insight into possible relations between accent types and visual information, it would appear necessary to annotate the recorded data in terms of existing labelling schemes like ToBI (Beckman & Ayers Elam, 1997) or ToDI (Gussenhoven, 2005), provided that such more detailed specifications of contours are reliable and reproducible.

Of course, the first study was limited in a number of ways. First of all, it was dealing with only one aspect of a newsreader's expressive style, as it focused on markers of prominent information only. Obviously, facial expressions may signal other kinds of socially or linguistically relevant information as well, like a speaker's attitude or emotion. In addition, the first study was very much speaker-oriented in that it did not really test how facial markers were perceived by addressees, other than that labellers had used a perceptual procedure to annotate the data in terms of auditory and visual properties. This in itself does not entail that such facial markers are exploited when decoding incoming utterances. And finally, the newsreader data of the first study were constrained to recordings of readers who are primarily addressing a specific type of audience, namely adult viewers. It remains to be seen whether the expressive style of newsreaders varies if they would bring the news to other audiences, like children. Therefore, the second study was designed to gain insight into these additional factors.

3. Study 2

3.1. Goal

The goal of the second study was to find out whether newsreaders adapt their facial expressions to the seriousness of the topic they are talking about, and, if so, whether this adaptation differs for newsreaders addressing adults or children.

3.2. Method

3.2.1. Materials

The stimulus materials were selected from both the child journal and the journal for adults. From a database of about 100 recordings, 40 items (20 adult data, 20 child data) were selected that were initially thought to represent lighter or more serious topics. The former were items related to topics like a big regional festivity or the introduction of a new computer game, whereas the latter would refer to topics like the capture of 15 British marines in Iran, or the war in Iraq. The newsreaders of the adult news were Philip Freriks and Sacha de Boer, while those for the child news were Milouska Meulens and Pepijn Crone. Fig. 2 gives a representative still of one of the newsreaders analysed (Milouska Meulens), while reading the child news (see Fig. 1 for a newsreader of the adult news).

To double check whether the initial classification of kinds of topics as being "light" or "heavy" made sense, all the topics were presented on paper to 15 independent participants. They were given a list with all the items randomly ordered, presented as one sentence summaries. Their task was to indicate on a 7-point scale whether they thought an item was rather light (1) or very serious (7), or representing a score in between those two extremes. The ratings from all participants on all 40 items were analysed with a repeated-measures analysis of variance with audience (adult, child) and subject (light, serious) as independent factors, and the score on the 7-point scale as dependent variable. This analysis revealed a main effect of the subject ($F_{(1,14)} = 363.607, p < 0.001, \eta_p^2 = 0.963$) and of audience ($F_{(1,14)} = 25.224, p < 0.001, \eta_p^2 = 0.643$). The effects turned out to be due to the fact that serious topics were indeed rated as being more serious than the lighter ones (serious: 5.755 (0.110); light: 2.562 (0.145)), whereas topics in the child news were on average considered to be somewhat lighter than the topics for the adult news (children: 3.918 (0.095); adults: 4.399 (0.121)). In addition, there was a significant 2-way interaction between audience and subject ($F_{(1,14)} = 28.646, p < 0.001, \eta_p^2 = 0.672$), which can be explained by looking at Table 4. It shows that while the serious topics appear to be equally serious in both kinds of news, the lighter subjects are somewhat lighter in the child news data.

Fragments of the video-clips covering these different topics were then cut out of their original context. Those fragments were about 10 s long on average, the shortest being 8 s and the longest 12 s. Fragments were always cut in such a way that the onset and off-set coincided with a sentence boundary. Given that we were

Fig. 2. Representative still of one of the newsreaders (Milouska Meulens) we analysed, while she was reading the child-directed news.

Table 4
Average scores (and errors) and difference scores of rated seriousness of light and serious news items (one-sentence written summaries) of newsreaders (adult and child news).

Audience	Topic		Δ−score (1–2)
	Light (1)	Serious (2)	
Children	2.107 (0.122)	5.730 (0.123)	−3.623
Adults	3.017 (0.200)	5.780 (0.109)	−2.763

exclusively interested in the non-verbal features of the newsreaders (especially their facial expressions) to the emotional content of their messages, we made sure that the clips themselves did not contain any other indicators to the news topic. In 18 clips, however, there were some pictures or movie clips displayed behind the newsreader which revealed what they were talking about. In these cases, we digitally covered those backgrounds images with a black square, so that the participants could only focus on the newsreader to deduce the emotional content of the news item.

3.2.2. Procedure

The experiment was an individually performed perception test, in which participants were presented with the selected clips in vision-only format. They saw the clips on a laptop computer, and had to give their scores on an answering sheet. To compensate for possible order effects, the clips were presented in two random orders, with half of the subjects seeing the clips in one order, and the other half in the opposite order. After having seen a clip, the participants had 5 s to produce two scores (on two separate 7-point Likert-scales), i.e., a first score where they had to indicate how expressive the newsreader was, with "1" meaning "not expressive" and "7" meaning "very expressive"; on the other scale, they had to express how serious they thought the topic was that the newsreader was talking about, with "1" meaning "not serious" and "7" meaning "very serious". The actual experiment was preceded by a short trial experiment with the same set-up, in which three practice fragments, comparable to the actual stimuli, but not used in the main experiment, were presented to the participants.

3.2.3. Participants

Fifty-four participants (19 men, 35 women) took part in the perception experiment with an average age of 30.5 years. None of them had participated in the first experiment.

3.3. Results

Both the data for expressivity and judged seriousness of the news items were analysed with a repeated measurements analysis of variance with topic (serious, light) and audience (adults, children) as within-subject factors, and gender of participant (male, female) as between-subject factor. We ran two analyses, one with the expressivity scores as dependent variable, one with the seriousness scores as dependent variable.

Expressivity: The repeated measurements anova with the expressiveness scores as dependent variable revealed a main effect of topic ($F_{(1,52)} = 158.613, p < 0.001, \eta_p^2 = 0.753$) and audience ($F_{(1,52)} = 135.657, p < 0.001, \eta_p^2 = 0.723$), while the effect of gender was not significant. Newsreaders were perceived as less expressive when bringing a serious topic than when talking about a light topic (serious topic: 4.109 (0.075); light topic: 4.874 (0.066)). In addition, newsreaders addressing children were on average more expressive than newsreaders addressing adults (children: 4.997 (0.061); adults: 3.986 (0.090)). In addition, there turned out to be a significant 2-way interaction between audience and topic ($F_{(1,52)} = 6.657, p < 0.05, \eta_p^2 = 0.013$), with all other interactions not significant. The interaction can be explained by looking at Table 5. It reveals that, while newsreaders addressing children are comparatively more expressive than adult newsreaders for both light and serious topics, the difference in expressiveness between those two kinds of topics is somewhat bigger for the newsreaders addressing children.

Table 5
Average scores (and errors) and difference scores of perceived expressiveness of newsreaders (adult and child news) in light and serious news items.

Audience	Topic		Δ−score (1–2)
	Light (1)	Serious (2)	
Children	5.455 (0.072)	4.539 (0.078)	0.916
Adults	4.293 (0.092)	3.679 (0.105)	0.614

Table 6
Average scores (and errors) and difference scores of perceived seriousness of newsreaders (adult and child news) in light and serious news items.

Audience	Topic		Δ−score (1–2)
	Light (1)	Serious (2)	
Children	3.071 (0.114)	4.331 (0.088)	−1.260
Adults	3.896 (0.121)	4.840 (0.097)	−0.944

Seriousness: When looking at the scores for seriousness, the repeated measurements anova similarly revealed a main effect of topic ($F_{(1,52)} = 142.205, p < 0.001, \eta_p^2 = 0.732$) and audience ($F_{(1,52)} = 47.074, p < 0.001, \eta_p^2 = 0.475$), while the effect of gender was again not significant. Newsreaders were perceived as being more serious when talking about a serious topic than about a light topic (serious topic: 4.585 (0.077); light topic: 3.484 (0.097)). The effect of audience was such that newsreaders addressing children were on average less serious (irrespective of the topic) than newsreaders addressing adults (children: 3.701 (0.077); adults: 4.368 (0.099)). As with the expressiveness scores, there was a significant 2-way interaction between audience and topic ($F_{(1,52)} = 5.319, p < 0.05, \eta_p^2 = 0.093$), with all other interactions again not being significant. Table 6 shows that the difference in perceived seriousness is bigger in the child data.

In order to gain some insight into the relation between the scores for expressiveness and seriousness, we measured correlations between the averages per topic for these two scores, which gave significant negative correlations for the child data ($r = -0.748, p < 0.01$), for the adult data ($r = -0.606, p < 0.01$), and for all data collapsed together ($r = -0.750, p < 0.01$). In other words, it appears that more serious topics are produced in a less expressive way, which trend is stronger for the child than for the adult data.

3.4. Discussion

The second study revealed overall that the emotional content of a message is reflected in the expressive style of newsreaders. Dependent on whether they report about a light or more serious topic, their visual appearance becomes more positive or negative. This effect interacts with that of another factor, i.e., whether newsreaders are addressing an adult or child audience: the former

appear to be less clear about the emotional content than the latter. For both kinds of speakers it holds that expressiveness correlates with the valency of the message with positive items being brought in a more expressive style than the negative ones.

Of course, the effects are subtle, as is clear from comparing the ratings from the text-only and video-only data. In the former, where participants had access to the actual content of the message (but could not see the newsreaders), the judged differences between serious and light messages are very marked, with difference scores of −3.623 for child data and −2.763 for adult data (see Table 4). The difference scores based on the perceptual ratings of the newsreaders' visual appearance for expressiveness and seriousness (see Tables 5 and 6) are far less extreme. This result in itself may not be surprising, given the conventions of the genre, as newsreaders are supposed to remain neutral and objective with respect to the topic at hand. As a matter of fact, some of them even explicitly maintain that their visual appearance does not reveal any cue about the emotional content of their news items.[2] In view of these claims, it is interesting to notice that, despite this objective to remain visually neutral, their facial expressions do leak some emotional content, even though, especially regarding the adult news items, there appears to be a bias toward "seriousness" overall, as all judgements are above 3.5 on the seriousness scale.

Even when we have tried to reduce the effect of possible confounding factors as much as possible, there are at least two issues that may need further analyses in the future. At this stage, the more extreme difference scores from the child data could be due to the more expressive differences between newsreaders or to the fact that the topics themselves were more different in content in the child data. First, the newsreaders of the adult news and the ones for the child news are of course different speakers. Therefore, it cannot be excluded—in principle—that the differences in expressive style and in the extent to which emotional content is revealed between the child and adult data are due to the fact that different speakers have produced those two datasets. That is, possibly, the differences might be related to idiosyncratic differences between extravert or introvert speakers who happen to be responsible for child and adult news, respectively. If so, the differences are not so much a matter of audience design or adaptation, but rather reflect differences in personality of the newsreaders (even when it would then still be interesting to explore why exactly these people were selected to read either adult or child news). Along the same lines, as displayed in Table 4, the news items of the adult and child data are not entirely comparable in terms of their intrinsic emotional content (as judged from text-alone judgments). While the serious topics in both data get near-identical ratings (5.730 for child and 5.780 for adult data), the scores for the selected light topics diverge somewhat more (2.107 versus 3.017, respectively). Unfortunately, given that we have based our analyses on found data and due to the nature of the data, the effects of speaker-specific and item-specific factors cannot easily be measured. Moreover, the newsreaders are public figures and may be known as such to the participants of our judgment experiments, so that it remains to be seen whether this knowledge has had an effect on their ratings.

4. General discussion

This study revealed that professional Dutch newsreaders make use of their expressive style for different communicative purposes. Subtle variations in their visual appearance appear to be used to support the information structure of their messages in that words which are prosodically marked as prominent tend to be accompanied by movements of the head and the eyebrows, in line with Cavé et al. (1996), Flecha-Garcia (2006a, 2006b, 2007), Graf et al. (2002), Yasinnik et al. (2004), Dohen (2005), Dohen and Lœvenbruck (2009), Dohen et al. (2004), Swerts and Krahmer (2008), Krahmer and Swerts (2007) and Scarborough et al. (2009). Our work reveals that such patterns also show up in more formal newsreader data; moreover, we found that pitch accents are accompanied by combinations of visual cues (head nods and eyebrows). In addition, the expressions of the newsreaders reveal the emotional content of the topic they address (i.e., whether it represents a light or serious item); the extent to which such connotations are signalled appear to be dependent on the intended audience, whereby child-directed news readings are more expressive than those in which adults are addressed. The latter outcome is consistent with previous findings that speakers adapt their non-verbal behaviour to characteristics of their addressees (Burnham et al., 2002; Hirschberg et al., 2004; Oviatt et al., 1999; Barkhuysen et al., 2005).

Such results indicate that newsreaders' data could indeed be a useful resource for studies of expressive style. In the past, such data have already been used as a basis for models of speech production or perception. For instance, the Boston radio news corpus (Ostendorf, Price, & Shattuck-Hufnagel, 1995) has been labelled in terms of the ToBI framework, and used a resource for prosodic modelling. Such analyses have both advantages and potential drawbacks. On the one hand, newsreader data are interesting for research purposes in that they represent a natural language domain, yet are sufficiently controlled to allow for specific hypothesis testing. In that sense, the setting resembles a laboratory situation in which possible confounding factors (specific situational or contextual variables) are excluded to a large extent. On the other hand, it remains to be seen how representative data gathered from newsreaders are for other kinds of discourse, e.g., discourse produced by non-professional speakers in more spontaneous interactions. Newscasters are generally trained to exhibit socially acceptable emotions, which often could deviate from natural emotion. For example, they are never expected to show emotions such as frustration, annoyance, depression, overjoyed, etc. Another issue is that newsreaders may use a somewhat artificial newsreader style, an "in-the-air" voice, which they may not use when speaking in other settings. In addition, they may even exhibit features, while reading the news, which are atypical: Ostendorf et al., for instance, report that their newsreaders have a tendency to accent every second word in their utterances, which is much more frequent than the average proportion of accented words in other types of discourse. Yet, at the same time, newsreaders are often treated as the "best" speakers within their respective language communities, so that it stands to reason that they are also treated as role models for theoretical and computational approaches to speech production and perception, both by humans and machines.

The results presented above are interesting in view of existing models of spoken communication. First, the visual correlates of information structure and emotional aspects reveal that newsreaders strive to make their expressions congruent with the content of what they say. If a word is important or emotionally loaded, then the face expresses those aspects. For markers of prominence, this has previously been argued to be a result of a preference of speaker to align auditory features with visual ones. The outcome that newsreaders also reveal the emotional content of their message visually through their expressive style, is surprising in view of the fact that newsreaders themselves often argue that they remain completely neutral with respect of the content of the message, in line with a general policy to try and be

[2] Sacha de Boer, personal communication.

as objective as possible about the content of their news items. The fact that the emotional content is nevertheless visible, suggests that it is almost an automatic result of a natural tendency to make facial expressions consistent with content. Second, the differences between the child and adult newsreader data fits nicely in current theories of communication which give a central role to adaptation, where characteristics of the addressee are explicitly taken into account in the forms of language use. Whereas previous research very much focused on alignment in terms of choice of words, syntactic structures or pronunciation, the current analysis brought to light that it may also occur on the level of expressive style.

Given these results, it is worthwhile to reflect on how useful such newsreader data could be as a resource for more application-oriented models for speech synthesis and recognition. As discussed above, in the past, newsreaders were often used as models for building synthetic speech systems, with the goal of making the resulting speech of such systems resemble their style as closely as possible (e.g., diphones and intonation patterns were often derived from the newsreader data). There is currently a growing interest in the development of synthetic versions of talking heads, that may function as new interfaces in human–machine interactions. However, while a lot of results have been achieved into the development of realistic faces, most of that research has focused on making the lip movements as realistic as possible to increase the intelligibility. There is still a lot that needs to be done to make the faces also natural in terms of their expressions. Second, for speech recognition, facial information could be used as an additional resource for improving the natural language component or dialogue manager of a spoken dialogue system. Previous work has shown that facial information may improve automatic speech recognition (Petajan, 1985), or can help to detect turns in an interaction where users become frustrated about the performance of a system (Barkhuysen et al., 2005; Wang et al., 2007). Along the same lines, the current analyses suggest that such visual information could also be useful to locate stretches of speech that are important to the discourse, or to derive the emotional content of a message. The crucial issue for making such datadriven analyses for speech synthesis and recognition realistic, is to have large datasets, that are marked or annotated in terms of facial expressions, so that specific machine-learning algorithms can be applied to find associations between facial variation and other kinds of information (like aspects of the discourse structure). For this purpose, it would be useful to have automatic and reliable procedures to label facial expressions in video-recordings of speakers.

There are a number of research problems that are worth exploring in the future. First, the data that have been investigated in the current experiments were always collected from newsreaders who were presenting the news alone, which is usually the situation on Dutch public TV. This is different from what happens on other channels where the news is brought by multiple speakers who are co-present and present in turn particular items. For a variety of reasons, this other situation may have an effect on the expressive style of readers. First, a presentation by two or more speakers seems to resemble a natural interaction more than the single presentations, given that the former allows for immediate feedback between speakers which possibly may lead to a more expressive style. Second, related to the previous point, it has been shown before that the communicative style of speakers may depend on presence effects. This was supported by the findings of an earlier study on emotional correlates of losing and winning situations in games, played by children who were doing this alone or together with a classmate (Shahid, Krahmer, & Swerts, 2008). It turned out to be the case that children were more expressive about winning or losing situations when together than when alone. Similarly, in a study on gesturing (Mol, Krahmer, Maes, & Swerts, 2009), it has been found that speakers gesture more frequently and with a larger amplitude when they describe a particular cartoon to a human speaker, as opposed to describing a story to a computer. Finally, our study has focused on newsreaders on Dutch public TV, which is generally considered to be a more "serious" channel, as it tries to bring news about important items in a neutral and objective way. This is potentially different for more commercial channels, that also concentrate on lighter items (related to show business and sports), which in turn may have an effect on the way these are being presented. Indeed, newsreaders may become more engaged when they are covering a topic like a sports event, which tends to induce more expressive behaviour.

Acknowledgements

This research was conducted within the FOAP project, which is funded by the Netherlands Organisation of Scientific Research (NWO) (see http://foap.uvt.nl). The second author also acknowledges NWO for Grant 277-70-007. We thank Ruud Koolen, Iris Pfrommer and Annette Bout for help with the data collection and visual annotations, and Lennard van de Laar for technical assistance. We thank the Dutch Broadcasting Foundation (NOS) for giving us permission to show the video stills of the newsreaders in Figs. 1 and 2. Finally, we would like to acknowledge Robert Dale, Alice Turk and two anonymous reviewers for valuable comments on an earlier version of this article.

References

Barkhuysen, P. N., Krahmer, E., & Swerts, M. (2005). Problem detection in human–machine interactions based on facial expressions of users. *Speech Communication*, *45*, 343–359.

Beckman, M., & Ayers Elam, G. (1997). *Guidelines for ToBI labelling*. The Ohio State University Research Foundation, Ohio State University.

Beskow, J., Granström, B., & House, D. (2006). Visual correlates to prominence in several expressive modes. In *Proceedings of interspeech*, Pittsburgh, PA (pp. 1272–1275).

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new pussycat? On talking to babies and animals. *Science*, *296*, 1435.

Cassell, J., Vihjálmsson, H., & Bickmore, T. (2001). BEAT: The behavior expression animation toolkit. In *Proceedings of SIGGRAPH'01* (pp. 477–486).

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 ICSLP*, Philadelphia (pp. 2175–2179).

Chafe, W. L. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. N. Li (Ed.), *Subject and topic*. New York: Academic Press.

Clark, H. H. (1992). *Arenas of language use*. Chicago: University of Chicago Press.

Dohen, M. (2005). *Deixis prosodique multisensorielle: Production et perception audiovisuelle de la focalisation contrastive on français*. Ph.D. thesis, Institut National Polytechnique de Grenoble (INPG), Grenoble.

Dohen, M., & Lœvenbruck, H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech*, *52*, 177–206.

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.

Flecha-Garcia, M. L. (2006a). Eyebrow raising, discourse structure, and utterance function in face-to-face dialogue. In *Proceedings of CogSci-2006*, Vancouver, Canada (pp. 1311–1316).

Flecha-Garcia, M. L. (2006b). *Eyebrow raising in dialogue*: *Discourse structure, utterance function, and pitch accents*. Ph.D. thesis, University of Edinburgh, UK.

Flecha-Garcia, M. L. (2007). Non-verbal communication in dialogue: Alignment between eyebrow raises and pitch accents in English. In *Proceedings of CogSci-2007*, Austin, TX, USA (p. 1753).

Graf, H. P., Cosatto, E., Strom, V., & Huang, F. J. (2002). Visual prosody: Facial movements accompanying speech. In *Proceedings of the fifth IEEE international conference on automatic face and gesture recognition* (*FGR 02*) (pp. 396–401).

Granström, B., House, D., & Lundeberg, M. (1999). Prosodic cues to multimodal speech perception. In *Proceedings of the 14th ICPhS*, San Francisco.

Gussenhoven, C. (2005). Transcription of dutch intonation. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 118–145). Oxford: Oxford University Press.

Hirschberg, J., Litman, D., & Swerts, M. (2004). Prosodic and other cues to speech recognition failures. *Speech Communication*, *43*, 155–175.

Jordan, T., & Sergeant, P. (2000). Effects of distance on visual and audio-visual speech recognition. *Language and Speech*, *43*(1), 107–124.

Krahmer, E., Ruttkay, Zs., Swerts, M., & Wesselink, W. (2002). Pitch, eyebrows, and the perception of focus. In *Proceedings of the speech prosody 2002*, Aix-en-Provence (pp. 443–446).

Krahmer, E., & Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech Communication*, *34*(4), 391–405.

Krahmer, E., & Swerts, M. (2004). More about brows: A cross-linguistic study via analysis-by-synthesis. In Zs. Ruttkay, & C. Pelachaud (Eds.), *Evaluating ECAs* (pp. 191–216). Dordrecht: Kluwer Academic Press.

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*, 396–414.

Ladd, D. R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

Mo, Y., Cole, J., & Lee, E.-K. (2008). Naïve listeners prominence and boundary perception. In *Proceedings of the speech prosody conference*, Campinas, Brazil.

Mol, L., Krahmer, E. J., Maes, A. A., & Swerts, M. (2009). The communicative import of gestures: Evidence from a comparative analysis of human–human and human–machine interactions. *Gesture*, *9*(1), 97–126.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility. Head movement improves auditory speech perception. *Psychological Science*, *15*, 133–137.

Ostendorf, M., Price, P. J., & Shattuck-Hufnagel, S. (1995). The Boston University radio news corpus. Boston University Technical Report No. ECS-95-001.

Oviatt, S. L., Bernard, J., & Levow, G. (1999). Linguistic adaptation during error resolution with spoken and multimodal systems. *Language and Speech*, *41*, 415–438.

Pelachaud, C., Badler, N., & Steedman, M. (1996). Generating facial expressions for speech. *Cognitive Science*, *20*, 1–46.

Petajan, E. (1985). Automatic Lipreading to Enhance Speech Recognition. In *Proceeding of the IEEE conference on computer vision and pattern recognition* (pp. 40–47).

Scarborough, R., Keating, P., Mattys, S. L., Cho, T., Alwan, A., & Auer, E. T. (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. *Language and Speech*, *52*, 135–175.

Shahid, S., Krahmer, E. J., & Swerts, M. (2008). Alone or together: Exploring the effect of physical co-presence on the emotional expressions of game playing children across cultures. In P. Markopoulus (Ed.), *Fun and games. Lecture notes in computer science*, Vol. 5294 (pp. 94–105). Berlin: Springer.

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

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America*, *101*(1), 514–521.

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

Wang, S., Demirdjian, D., Kjellström, H., & Darrell, T. (2007). Multimodal communication error detection for driver-car interaction. In *Proceedings of ICINCO 2007*, Angers, France.

Wilting, J., Krahmer, E., & Swerts, M. (2006). Real vs. acted emotional speech. In *Proceedings of the international conference on spoken language processing* (*interspeech 2006*), Pittsburgh, PA, USA, September 2006.

Yasinnik, Y., Renwick, M., & Shattuck-Hufnagel, S. (2004). The timing of speech-accompanying gestures with respect to prosody. In *Proceedings from sound to sense: 50+ years of discoveries in speech communication* (pp. C97–C102). Cambridge, MA: MIT.