Congruent and incongruent audiovisual cues to prominence

Marc Swerts & Emiel Krahmer

Communication & Cognition
Tilburg University, The Netherlands
{m.g.j.swerts;e.j.krahmer}@uvt.nl

Abstract

The current paper addresses the effect of auditory and visual information on the perception of accents. The research consists of two perception experiments in which we present video clips of recorded speakers as stimuli to listeners. The first experiment tests whether listeners can detect the accented syllable in a sequence of three nonsense syllables, which are presented to subjects in three conditions (audio+vision, vision alone, audio alone). The second experiment exploits so-called mixed stimuli, i.e., artificially constructed three-syllable utterances that have conflicting auditory and visual cues to accents. Results from these two studies confirm earlier findings that there are indeed visual cues to accents, but these appear to have weaker cue value than auditory information.

1. Introduction

There are various kinds of evidence showing that visual cues from a speaker's face have an impact on how a listener processes incoming speech. In particular, it appears that facial information can both support and interfere with purely auditory cues. On the one hand, a talking head, be it human or synthetic, can facilitate the speech decoding process: e.g. utterances become more clearly understandable when a listener can also see the face of the person who is talking (lipreading support) (Agelfors et al. 1998; Benoit et al. 2000). On the other hand, visual cues can compete with auditory ones. This is most clearly demonstrated by the well-known McGurk effect, where the perception of a particular CV-sequence changes when a listener is presented with conflicting visual information (e.g. /ba/ becomes /da/ when /ga/ is visually presented). Along the same lines, Pourtois et al. (2002) showed that listeners find it more difficult to process words spoken with a certain emotional tone (e.g. happy), when they are simultaneously looking at a face that expresses an incongruent emotion (e.g. sad). Such experiments with conflicting cues have been carried out to learn more about crossmodal perception and about the relative cue strength of auditory and visual features.

The current paper addresses the effect of auditory and visual information on the perception of accents, i.e., prominent syllables in a spoken utterance. While it is well known that speakers use verbal features, such as intonation, loudness and vowel lengthening, to highlight particular parts of their utterances, there is growing evidence that they mark them by visual cues as well. Following earlier claims by Ekman (1979), various people have suggested that eyebrow movements can signal prominent words in an utterance. They are therefore also modelled as markers of accents in embodied conversational agents (e.g. Cassell et al. 2001), though there is no consensus on their timing or placement. Empiricial evidence for these earlier claims comes from Keating et al. (2003) and Cavé et al. (1996), who report correlations between accented words and eyebrow movements, especially in the left eyebrow. Yet, there is no 1-to-1 relationship between the two, as not all accents are accompanied by eyebrow movements. Moreover, they appear to be more typical as markers of phrasal accents rather than of word stress. This outcome, which suggests that auditory markers of accents are relatively more important than eyebrow movements, is in line with perceptual results: listener evaluations, either with real (Keating et al. 2003) or synthetic talking heads (Krahmer et al. 2002a, b) as stimuli, reveal a rather modest contribution of eyebrow movements on the perception of accents. Of course, in addition to eyebrow movements, there may be other facial gestures as well that can function as visual markers of accents. Apart from clearly visible head nods, important cues may for instance be located in the mouth area of the face. Keating et al. (2003) found that some of their speakers produce accented words with greater interlip distance and more chin displacement. Similarly, Erickson et al. (1998) showed that the increased articulatory effort for realizing accented words correlates with more pronounced jaw movements.

The study reported here builds on our earlier studies (Krahmer et al. 2002a, 2002b), in which we also focussed on the combination of auditory and visual cues for signaling accents, but is new in two important respects. First, rather than using an analysis-by-synthesis method with parametricized synthetic faces, we now use an analysis-by-observation technique, where we explore real human speakers whose utterance productions are filmed. Second, while our previous studies limited the visual analyses to eyebrow movements only, we currently investigate possible cues in the face as a whole, given the potential importance of other features, e.g. in the mouth area. Our study consists of two perception experiments in which we present video clips of recorded speakers as stimuli to listeners. The first experiment tests whether listeners can detect the accented syllable in a sequence of three nonsense syllables, which are presented to subjects in three conditions (audio+vision, vision alone, audio alone). The second experiment exploits so-called mixed stimuli, i.e., artificially constructed three-syllable utterances that have conflicting auditory and visual cues to accents. Results from these two studies confirm earlier findings that there are indeed visual cues to accents, but these appear to have weaker cue value than auditory information.

2. Experiment 1: Congruent stimuli

2.1. Stimuli

Twenty native speakers of Dutch, colleagues and students from Tilburg University, volunteered as subjects in a brief speech elicitation task. Speakers were given small cards (each subject got a different random order) on which three CV syllables were

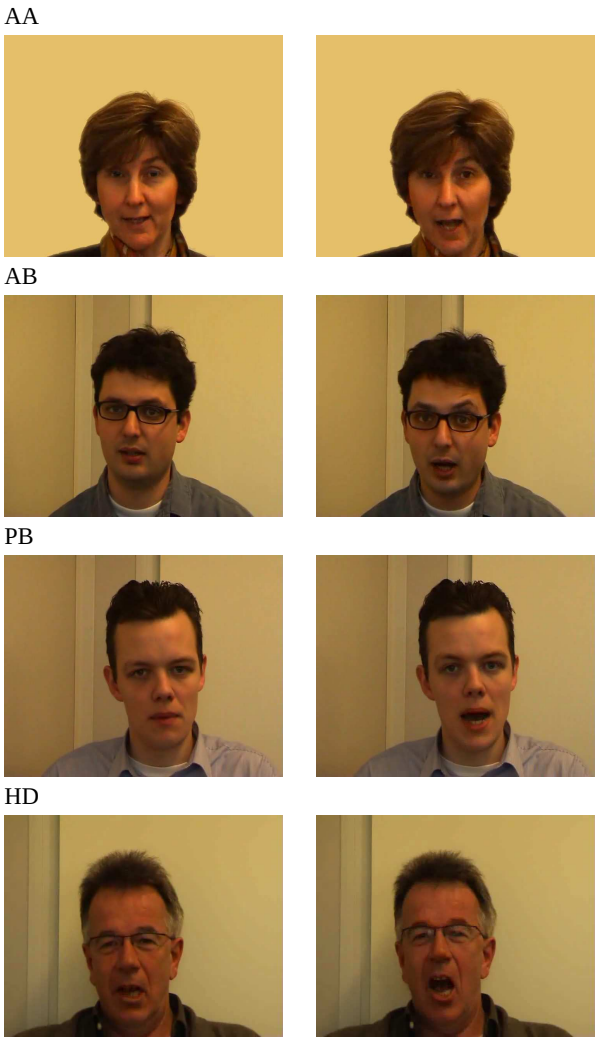

Figure 1: *Eight stills from recordings of four of our speakers (AA, AB, PB, HD) extracted from their production of a syllable in unaccented (left) and accented (right) condition.*

printed, either "ma ma ma" or "ga ga ga", where one of the three syllables was capitalized. The reason to use both labial and guttural consonants was to see whether frontal sounds would have clearer visual correlates of prominence than sounds produced in the back. The speakers were given the task to utter these syllables such that the one which was capitalized sounded more prominent than the other two. While speaking, they were being filmed (frontal view of the head). During a few test trials, the experimenter gave some indication as to the optimal tempo of the speech utterances, in order to get a normal speaking rate. Speakers had to utter the three-syllable sequences in two different versions, a neutral speaking mode versus an exaggerated one, where speakers had to imagine that they were addressing someone standing at a larger distance. Out of these different versions, we retained recordings of five speakers to be used as stimulus materials for perceptual testing. Their data were selected on the basis of a few criteria: whether they looked into the camera, whether their speech was not too variable regarding speech tempo, whether their speech was clearly audible, and whether they were at all able to produce the intended accents. Comparable to Keating et al. (2003), we found that some speakers had difficulty with the task to highlight a particular syllable. Figure 1 gives eight stills from recordings of four of our speakers extracted from their production of a syllable in unaccented (left) and accented (right) condition.

2.2. Procedure

45 listeners, students from the universities of Tilburg and Eindhoven (all different from the original speakers who produced the stimuli) participated as subjects in the current perception experiment. One third of the subjects was presented with the original videoclips (vision+sound), another third only heard the speech without seeing the speaker (sound only), and the final third saw the clips, but without the speech (vision only). In all tasks, subjects were asked to indicate — by forced choice — which of the three syllables they heard or saw was spoken with the strongest accent. Each stimulus was presented twice to the subjects; they were told to make a choice after the first presentation, which they could verify and possibly correct after the second presentation. The actual experiment was preceded by a short trial session to make people acquainted with the experimental procedure and the types of stimuli. No feedback was given on the "correctness" of their responses and there was no further interaction with the experimenter. Each condition of the total experiment lasted about 15 minutes.

2.3. Results and discussion

Table 1 gives overall classification results (not broken down into different conditions) for the three experiments, expressed as three 3-by-3 confusion matrices. A multinomial logistic regression analysis with accent position (first, second, third), modality (audio+vision, vision alone, audio alone), speaking condition (normal, exaggerated), speaker (5 speakers), and type of nonsense-syllable (ma, ga) as independent factors, and perceived accent as dependent variable, revealed main effects of 4 factors (accent position: χ^2=4931.678, df=4, p$<$.001; modality: χ^2=16.931, df=4, p$<$.01; speaking condition: χ^2=9.867, df=2, p$<$.01; speaker: χ^2=36.265, df=8, p$<$.001), and no significant effect for type of consonant-vowel sequence. In addition, there were significant 2-way interactions (all p$<$.05) between accent and modality, between modality and speaker, between speaker and nonsense-syllable and between nonsense-syllable and modality. When expressed in terms of percentage correct, the overall scores correspond with 97.11% correct classification for vision+sound stimuli, 97.33% for sound-only stimuli and 92.89% for vision-only stimuli. These scores indicate that the perception task was a fairly easy one, in all three conditions, though the vision-only stimuli are significantly less accurately classified as the other two conditions. Yet, since there is a clear ceiling effect, it is difficult to establish the relative cue strength of the auditory and visual cues. Therefore, we set up a second experiment, in which we deliberately manipulated our original stimuli such that they contained incongruent auditory and visual cues to accents.

3. Experiment 2: Incongruent stimuli

3.1. Stimuli

In this experiment, use was made of so-called mixed stimuli, i.e., stimuli artificially created from the original nonsense utterances using the Adobe Premiere video editing software pack-

Table 1: *Classification results of first, second and third accents in three experimental conditions: vision+speech. speech only, vision only*

Experimental Condition	Produced Accent	Perceived accent 1	2	3	Total
Vision+Sound	1	292	8	0	300
	2	10	286	4	300
	3	4	0	296	300
Total		306	294	300	900
Sound only	1	288	10	1	299
	2	3	293	4	300
	3	4	2	294	300
Total		295	305	299	899*
Vision only	1	285	14	1	300
	2	22	274	4	300
	3	6	17	277	300
Total		313	305	282	900

* one missing value

age. The newly created stimuli had conflicting visual and auditory cues to accent position in that the two never occurred on the same syllable. In other words, this experiment only contained stimuli with a mismatch between the visual and auditory cues. As a basis for the generation of these stimuli, we took the utterance materials from two speakers (AB and PB) whose data had most accurately been classified in the first experiment so that we only included cases with clear auditory and visual cues in the current experiment. In addition, we recorded one other speaker (LL) whom was asked to exaggerate the facial expressions. The reason to have one extra recording was to test whether there is evidence for a gradient effect of visual cues on perception of accents. The different syllables from the selected speakers were similar in length which was of importance to be able to allow for crossmodal editing. In order to make sure that the mismatch in alignment between the auditory and visual cues was not too strong, we had to insert some small pauses in the waveform in a small set of the cases, so that the auditory and visual information became properly aligned. This manipulation, however, did not affect the naturalness of the auditory signal nor the original perception of the accent in the speech only condition.

3.2. Procedure

55 listeners (students from the universities of Tilburg and Eindhoven), different from the original speakers, participated in the listening experiment. They were presented with all the mixed stimuli in a random order, and were instructed to indicate -by forced choice- which of the three syllables they observed was spoken with the strongest accent. As in the previous test, each stimulus was presented twice to the subjects; they were again told to make a choice after the first presentation, which they could verify and possibly correct after the second presentation. The actual experiment was again preceded by a short trial session to make people acquainted with the experimental procedure and the types of stimuli. No feedback was given on the "correctness" of their responses and there was no further interaction with the experimenter. The total experiment lasted about 10 minutes.

3.3. Results and discussion

Table 2 presents the classifications results, expressed as three 2-by-3 confusion matrices, for the utterances of our three speakers. A multinomial logistic regression analysis with auditory accent (first, second, third), visual accent (first, second, third) and speaker (AB, PB, LL) as independent factor, and perceived accent as dependent variable, revealed main effects of all independent factors (auditory accent: χ^2=1048.112, df=4, p<.001; visual accent: χ^2=167.843, df=4, p<.001; speaker: χ^2=14.997, df=4, p<.01). In addition, there was a significant 2-way interaction between speaker and visual accent (χ^2=35.506, df=8, p<.001).

Table 2: *Classification results of first, second and third accents for each combination of an auditory and visual accent*

Auditory Accent	Visual Accent	Perceived accent 1	2	3	Total
1	2	153	10	2	165
1	3	154	1	10	165
2	1	36	128	1	165
2	3	1	153	11	165
3	1	61	6	98	165
3	2	2	38	125	165
Total		407	336	247	990

Table 2 shows that there is an effect of utterance position on the perceptibility of an accent. That is, the effect of auditory information is stronger for initial syllables and weaker for final syllables. This decrease in perception with syllable position may be due to a declination effect in that pitch accents that occur later in a phrase are known to be less pronounced than initial ones. These positional effects are different from those reported by Keating et al. (2003), who found that there was speaker-dependent variation regarding the strength of visual markers of first or second syllables in 2-syllable words, where some speakers produced stronger cues on the first syllable and others on the second.

Table 3: *Percentage of perceived accents on syllables that got an auditory accent, a visual accent or neither an auditory or visual accent*

Speaker	Scores Auditory	Visual	Neither
AB	86.7%	10.6%	2.7%
PB	83.3%	15.8%	0.9%
LL	75.8%	24.0%	0.2%

Table 3 shows that —overall— the auditory cues are stronger than the visual cues, though the latter cannot be ignored. Also, the perception of accents depends on the speaker who produced the utterances. That is, AB has relatively stronger auditory cues, whereas for the stimuli of LL, who was instructed to exaggerate the facial expressions, the auditory cues have less impact. In addition, if we only focus on the auditory cues, it is interesting to see that the overall performance clearly degrades with respect to the results of the speech-only condition of our first experiment, where correct classifications were in the nineteen. This is in line with comments we got from our subjects,

who complained that the test was relatively difficult, whereas we had not received such comments after the first test. Also, subjects had noticed some problems with the stimuli: they had experienced that the some stimuli had been manipulated while it was difficult to make clear what exactly was wrong. Note also that the results on accent perception differ from the outcome of studies testing the McGurk effect on phoneme perception. There, it is typically found that the presentation of consonant-vowel combinations with inconsistent visual and auditory properties leads to the perception of a sound which is different from both the original auditory and visual signal from which the experimental stimuli were created. Unlike those earlier results, our current findings on accent perception show that only a neglectible fraction of the incongruent stimuli have perceived accents that were neither cued by auditory or visual features.

4. General discussion and conclusion

As noted in our introduction, there is a whole body of research showing that speakers use verbal features, such as intonation, loudness and vowel lengthening, to highlight particular parts of their utterances. While our current study confirms the cue value of auditory features for the perception of accents, we have provided evidence that visual features are important as well. As a matter of fact, the facial expressions, while being less strong than the auditory cues, function surprisingly well as cues to accents. In our first experiment, both auditory and visual accents yield almost perfect classification results. However, due to these strong ceiling effects, it was difficult to establish the relative cue strength of auditory and visual features. Therefore, the second experiment uses mixed stimuli with auditory and visual cues that were incongruent as signals for accent. This test clearly showed that the speech cues are predominant, but that visual cues can interfere with auditory information, in attracting some of the perceived accents.

We intend to further explore how exactly visual accents are being encoded by speakers in their facial expressions, and which facial area is more important for accent perception. Results from Keating et al. (2003) suggest that speakers may differ regarding their use of visual cues. That is, some speakers more often use head nods, some vary features in the mouth area, while others exploit eyebrow movements, where there can even be a difference between the use of the left and the right eyebrow (see also Cavé et al. 1996). Therefore, we are currently setting up experiments in which we systematically blacken particular areas in a talking face (e.g. upper part versus lower part or left versus right part), to see which area has stronger impact on accent perception. Next, there were indications from our two experiments that the relative strength of visual cues was somewhat speaker-dependent, meaning that some speakers exploit visual cues more than others, or that they differ in the degree to which they exploit a particular parameter. Therefore, it might be interesting to study thresholds for combinations of visual and auditory cues. In line with previous studies, it is fairly straightforward to create pitch continua (e.g. gradual increase in excursion size of pitch movements) which are known to correspond with different perceived degrees of prominence (e.g. 't Hart et al. 1990). It would be nice to combine such continua with (artificially) induced visual continua as well, such as continua in eyebrow movement, articulatory movements or head nods. This could be done, either with synthetic heads or stimuli recorded from real speakers. In line with earlier studies by Pourtois et al. (2002), one could also investigate whether stimuli that are inconsistent regarding their use of visual and auditory cues to accent are more difficult to process than stimuli where the two types of cues do match. Finally, following Keating et al. (2003), one could study the same phenomena in real words and utterances. Using nonsense-syllables obviously has the advantage of yielding clear data, yet they may differ from data in naturalistic settings.

Acknowledgments This research was conducted as part of the VIDI project "Functions of Audiovisual Prosody (FOAP)", sponsored by the Netherlands Organization for Scientific Research (NWO). Marc Swerts is also affiliated with Antwerp University and with the Fund for Scientific Research - Flanders (FWO-Flanders). Thanks to Lennard van de Laar and Iris Boshouwers (Tilburg University) for their help in carrying out the different experiments.

5. References

[1] Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E, Öhman, T. 1998. Synthetic faces as a lipreading support. in: *Proc. ICSLP*, Sydney, Australia.

[2] Benoit, C., Martin, J.-C., Pelachaud, C., Schomaker, L. & Suhm, B. (2000). Audio-Visual and Multimodal Speech Systems, in: *Handbook of Standards and Resources for Spoken Language Systems*, D. Gibbon, I. Mertens, R. Moore (eds.), Kluwer Academic Publishers.

[3] Cassell, J., Vihjálmsson, H., Bickmore, T. 2001. BEAT: the Behavior Expression Animation Toolkit, *Proc. SIGGRAPH'01*, 477–486, Los Angeles.

[4] Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., Espesser, R. 1996. About the relationship between eyebrow movements and F_0 variations, *Proc. ICSLP*, Philadelphia, pp. 2175–2179.

[5] Ekman, P. 1979. About brows: Emotional and conversational signals, in: *Human ethology: Claims and limits of a new discipline*, M. von Cranach, K. Foppa, W. Lepenies, D. Ploog (eds.), Cambridge University Press. pp. 169–202.

[6] Erickson, D., Fujimura, O., Pardo, B. 1998. Articulatory correlates of prosodic control: Emotion versus emphasis, *Language and Speech; Special Issue on Prosody and Conversation*, Vol.41, No.3-4, 399–417.

[7] t Hart, J., Collier, R., Cohen, A. 1990. *A perceptual study of intonation.* Cambridge: Cambridge University Press.

[8] 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 *Proc. ICPhS* Barcelona, Spain, 2071–2074.

[9] Krahmer, E., Ruttkay, Zs., Swerts M., Wesselink, W. 2002a. Pitch, Eyebrows and the Perception of Focus. *Proc. Speech Prosody 2002*, Aix-en-Provence, France, April 2002.

[10] Krahmer, E., Ruttkay, Zs., Swerts M., Wesselink, W. 2002b. Audiovisual Cues to Prominence. In: *Proc. ICSLP*, Denver, USA, September 2002.

[11] Pourtois, G., Debatisse, D., Despland, P. A., de Gelder, B. 2002. Facial expressions modulate the time course of long latency auditory brain potentials. *Cognitive brain research* 14, 99–105.