The interplay between the auditory and visual modality for end-of-utterance detection

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

Published in:
Journal of the Acoustical Society of America

Publication date:
2008

The interplay between the auditory and visual modality for end-of-utterance detection

Pashiera Barkhuysen, Emiel Krahmer, and Marc Swerts[a)]
Communication & Cognition, Faculty of Arts, Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands

(Received 2 June 2006; revised 22 October 2007; accepted 26 October 2007)

The existence of auditory cues such as intonation, rhythm, and pausing that facilitate end-of-utterance detection is by now well established. It has been argued repeatedly that speakers may also employ visual cues to indicate that they are at the end of their utterance. This raises at least two questions, which are addressed in the current paper. First, which modalities do speakers use for signalling finality and nonfinality, and second, how sensitive are observers to these signals. Our goal is to investigate the relative contribution of three different conditions to end-of-utterance detection: the two unimodal ones, vision only and audio only, and their bimodal combination. Speaker utterances were collected via a novel semicontrolled production experiment, in which participants provided lists of words in an interview setting. The data thus collected were used in two perception experiments, which systematically compared responses to unimodal (audio only and vision only) and bimodal (audio-visual) stimuli. Experiment I is a reaction time experiment, which revealed that humans are significantly quicker in end-of-utterance detection when confronted with bimodal or audio-only stimuli, than for vision-only stimuli. No significant differences in reaction times were found between the bimodal and audio-only condition, and therefore a second experiment was conducted. Experiment II is a classification experiment, and showed that participants perform significantly better in the bimodal condition than in the two unimodal ones. Both the first and the second experiment revealed interesting differences between speakers in the various conditions, which indicates that some speakers are more expressive in the visual and others in the auditory modality. [DOI: 10.1121/1.2816561]

PACS number(s): 43.70.Mn, 43.71.Sy, 43.71.An, 43.71.Bp [ARB] Pages: 354–365

I. INTRODUCTION

Speakers use nonlexical features to demarcate various kinds of speech units, varying from a simple phrase to a larger scale discourse segment or a turn in a natural conversation. Previous studies have largely focused on how prosodic variables, such as intonation, rhythm and pause, or more subtle modulations of voice quality, like creaky voice, can be exploited to signal the end of such units (e.g., de Pijper and Sanderman, 1994; Price *et al.*, 1991; Swerts *et al.*, 1994a; Wightman *et al.*, 1992). In addition to features that are encoded in the speech signal itself, there is also an investigation into how particular visually observable variations from a speaker's face, like gaze patterns or bodily gestures, can be used as boundary cues (e.g., Argyle and Cook, 1976; Cassell *et al.*, 2001; Nakano *et al.*, 2003; Vertegaal *et al.*, 2000). However, little is known about the perception of these visual cues, and about the relative importance of the visual and the auditory modality for demarcation purposes. Therefore, the aim of this paper is to get more insight into which modalities speakers use for signaling finality or nonfinality, and how sensitive observers are to these respective signals. In particular, our goal is to investigate the relative contribution of three different conditions to end-of-utterance detection: two unimodal ones, vision only and audio only, and their bimodal combination.

It is by now well established that various auditory cues may serve as boundary markers of speech utterances (e.g., Koiso *et al.*, 1998; de Pijper and Sanderman, 1994; Price *et al.*, 1991; Swerts *et al.*, 1994a; Ward and Tsukahara, 2000; Wightman *et al.*, 1992, among many others). One of the strongest prosodic indicators for the end of a speaker's utterance is a pause, either a silent interval or a filler such as "uh" and "uhm" (as shown by, among others, de Pijper and Sanderman, 1994; Price *et al.*, 1991; Swerts, 1997, 1998; Wightman *et al.*, 1992). Many of these studies are based on analyses of monologues, where it was even found that pause length may covary with the strength of a boundary. When looking at natural interactions between multiple speakers, however, pauses tend to be rather short inbetween two consecutive speaker turns. Even though end-of-utterance pauses may be very short in interaction, turn switching proceeds remarkably smoothly, generally without overlap between speakers (Koiso *et al.*, 1998; Levinson, 1983; Ward and Tsukahara, 2000).

One of the reasons why the turn-taking mechanism may proceed so fluently, is that speakers "presignal" the end of their utterances (e.g., Couper-Kuhlen, 1993; Caspers, 1998; Swerts *et al.*, 1994a, Swerts *et al.*, 1994b). Listeners may pick up these cues and therefore may know in time when the current turn will be finished. Various researchers have looked

[a)]Author to whom correspondence should be addressed. Electronic mail: m.g.j.swerts@uvt.nl

in detail at the nature of these cues. It has been suggested, for instance, that the capacity of listeners to anticipate an upcoming boundary is based on what is called rhythmic expectancy (Couper-Kuhlen, 1993). Related to this, there is subtle durational variation, such as preboundary lengthening, which speakers can use to mark the final edge of a speech unit such as a turn (Wightman *et al.*, 1992; Price *et al.*, 1991). In addition to these timing-related phenomena, many researchers have focused on the potential use of melodic boundary markers as well. First, there are local boundary markers which occur at the extreme edge of a turn unit, right before an upcoming boundary, for which it has been shown that tones which reach a speaker's bottom range clearly function as finality cues (Swerts and Geluykens, 1994; Caspers, 1998; Koiso *et al.*, 1998). Moreover there appear to exist melodic structuring devices which are more global in nature in that they are spread over a whole speech unit. In particular, various studies have pointed out that speech melody gradually decreases in the course of an utterance, which may enable listeners to feel a boundary coming up (e.g., Leroy, 1984). However, this declination pattern may be typical of read-aloud speech which allows for a larger degree of look-ahead compared to spontaneous speech. Other finality cues are variations in pitch span, and more subtle differences in the alignment of pitch movements (Silverman and Pierrehumbert, 1990; Swerts, 1997). Finally, there is acoustic evidence which shows that marked deviations from normal phonation, in particular, creaky voice, typically occur at the end of an utterance (Carlson *et al.*, 2005).

The possible premonitoring cue value of prosodic cues has been explicitly tested in various perception studies. Grosjean (1983) and Leroy (1984) have already established that human subjects are surprisingly accurate in estimating the location of an upcoming boundary, using a variant of a gating paradigm, in which listeners are only presented with the initial part of an utterance. Along the same lines, Swerts *et al.* (1994a) and Swerts and Geluykens (1994) reported that people are able, on the basis of melodic cues, to judge the serial position of a phrase in a larger discourse unit. Carlson *et al.* (2005) found that native speakers of Swedish and of American English showed a remarkable similarity in judgments when they had to predict upcoming prosodic breaks in spontaneous Swedish speech, even when they had to base such estimations on stimuli that consisted of only a single word.

It thus seems safe to conclude that speakers and listeners take the auditory modality into account while marking the end of an utterance. But to what extent do they pay attention to the visual modality? Various researchers have argued that speakers may use visual cues for end-of-utterance signaling, where most studies have investigated how various bodily gestures may be used as markers of discourse boundaries. First, different studies focused on general changes in posture (Beattie *et al.*, 1982; Cassell *et al.*, 2001; Duncan, 1972). These studies suggest there is a general trend for people to change their pose when they start speaking, whereas they return to their initial posture at the end of a turn, for instance by raising their shoulders at the onset of a turn and lowering them again at the end. Second, one specific visual cue which has received much scholarly attention is related to movements of the eyes. Argyle and Cook (1976) describe in detail how the tuning of gaze behavior regulates many aspects of the interaction in a very subtle way. In general, it appears to be the case that speakers divert their gaze rather often while talking, whereas the listening conversation participant tends to look at the partner more frequently. When analyzing the gaze patterns in normal interactions more closely, it appears that a pattern emerges which is connected to the turn-taking mechanism, in that speakers tend to divert their gaze when they start talking, and return the gaze to their partner when they are finished (see also Goodwin, 1980; Kendon, 1967; Nakano *et al.*, 2003; Novick *et al.*, 1996; Vertegaal *et al.*, 2000). The cue value of gaze is likely to be due to the fact that human eyes have a unique morphology, with a large white sclera surrounding the dark iris. It has been argued that this contrast may have evolved to make it easier to detect the gaze direction of others (Kobayashi and Kohshima, 1997). While variation in posture shifts and gaze patterns have been directly linked to boundary marking, in particular in the turn-taking system, various researchers have argued that there may be further visual cues that may be important for demarcation purposes as well, such as head nods (e.g., Maynard, 1987), eyebrow movements (e.g., Ekman, 1979; Krahmer and Swerts, 2004), and eye blinks (e.g., Doughty, 2001).

The results from the various studies described above thus suggest that a speaker can display that he or she is going to stop speaking, by means of both auditory and visual features. However, there are still a large number of unsolved questions regarding the relative importance of the modalities and of their combined effects. While it has been shown that listeners are accurate in determining the end of an utterance based on the auditory modality, it is unknown whether they would be equally capable of doing so on the basis of visual information as well. And if so, it is still an empirical question as to how the visual modality relates to the auditory one, whether or not the two modalities may reinforce each other, and whether observers are helped or rather distracted when they have to focus on two rather than on a single modality in their finality judgments.

To this end, we have set up two experiments that are both based on perceptual judgments of stimuli in one of three conditions: a vision-only, audio-only, or audio-visual condition. The experiments make use of audio-visual recordings of semispontaneous utterances that were naturally elicited in a question-answering paradigm. The first experiment explores differences between modalities via a reaction time experiment in which participants are instructed to indicate as soon as possible when they think an utterance, presented in one of three conditions, ended. The second experiment makes use of basically the same stimuli as the ones from the first experiment, and looks in more detail at which factors influence participants' abilities to judge whether a speaker's turn is about to end or not; in this experiment, subjects are presented both with longer and shorter speech fragments, so we may get insight into the cue value of possible global versus local cues to finality. In addition, we look in more detail into the question of which auditory and visual cues are actually used by our speakers.

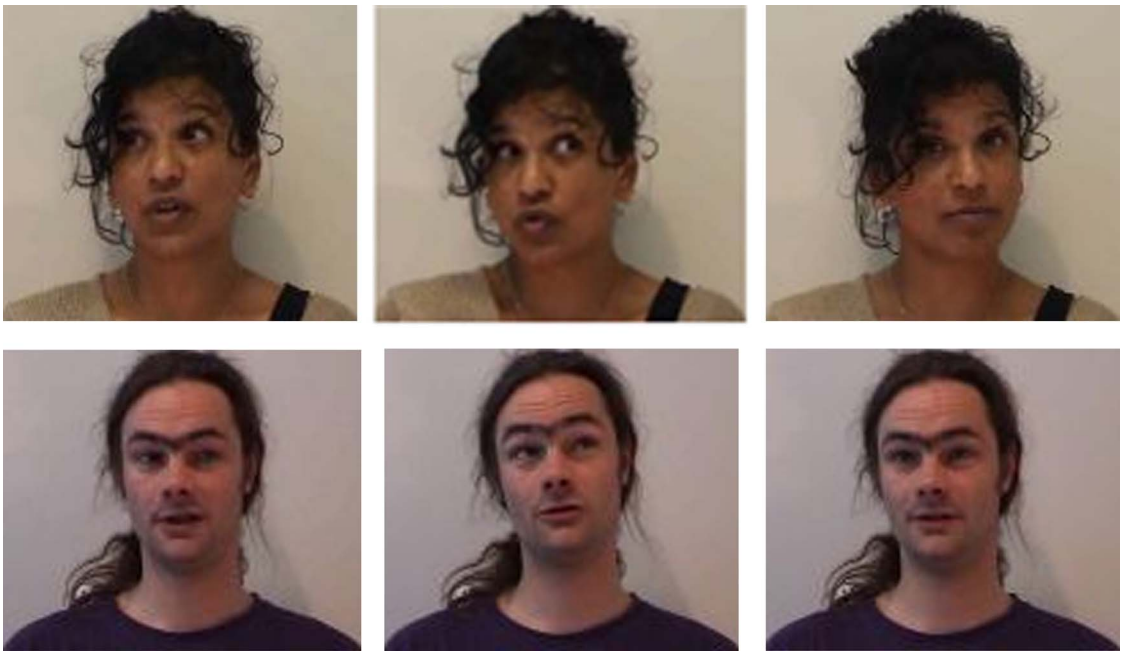

FIG. 1. (Color online) Representative stills of speakers SS (top) and BB (bottom) while uttering the first and middle word and just after uttering the final word of a three word answer, such as "red, white, blue."

II. AUDIO-VISUAL RECORDINGS

We gathered digital video recordings of speakers responding to questions in a natural, interview-style situation. Although recent research suggests that lexical and syntactic factors are relevant for end-of-utterance detection (de Ruiter *et al.*, 2006), for our current purposes, however, these factors should be eliminated as they would offer an unfair advantage to the auditory modality. Hence the questions were intended to elicit lists of words, where the lexical and syntactic structures of the answers offer no clues at all about where the end of the utterance is to be expected.

The questions were selected in such a way that they resulted in a variety of different answers, and such that potential answer words could occur in different positions in the list, depending on the question. Target answers varied in length, consisting of three or five words. Twelve questions were asked for predictable sets of numbers, in different orders, and with different number ranges. For instance: what are the multiples of five below 30?; what are the odd numbers below ten in reversed order?; and what are the multiples of five below 30 in reversed order?

Notice that the word "five" can occur both in a final and in a nonfinal position. The other questions addressed general knowledge or individual preferences of the interviewee, such as: what are the colors of the Dutch flag?; what are your three favorite colors?; and name five countries where you can go skiing.

Notice that for the second category the answers are never fully predictable. Even the colors of the Dutch flag are described by participants both as "red, white, blue" and "blue, white, red." Moreover, both "red" and "blue" can occur (and do in fact occur) as the second, middle word, in responses to the favorite color question. The interview consisted of 33 questions, of which 25 were experimental and eight were filler items. As filler items, questions were used for which the number of words in the answers could in principle not be predicted (e.g., Which languages do you speak?). These filler questions were added for the sake of variety and to make sure that speakers did not only produce three and five word lists.

A total of 22 speakers participated (13 male and nine female), between 21 and 51 years old. None of the speakers was involved with audio-visual research, and speakers did not know for what purpose the data were collected. The original recordings were made with a digital video camera [MiniDV; 25 frames/s, a resolution of 720×576 pixels, sampling of 4:2:0 (PAL), luma 8 bits chroma and 2 channel audio recording at 16 bits resolution and 48 kHz sampling rate]. The recordings were subsequently read into a computer and orthographically transcribed. See Fig. 1 for some representative stills.

III. EXPERIMENT I: REACTION TIMES

As a first exploration we performed a reaction time experiment with the intention to gain insight into the relative contribution of the auditory and visual modality, alone and in combination, for end-of-utterance detection.

A. Method

1. Stimuli

For this experiment four male and four female speakers were randomly selected from the corpus of 22 speakers described above. For each speaker, three instances of answers consisting of three words and three instances of five words were randomly selected on the basis of the transcriptions (8 speakers $\times$ 6 instances=48 stimuli in total). Notice that since this first selection was random, the set of selected an-

swers differed for each of the selected speakers. As a result, the lexical content of the selected answer lists was highly varied, and since words could occur in various (final and nonfinal) positions, observers could never rely on lexical information for their end-of-utterance detection. If the first selection contained answers with more than just list words (e.g., repetitions of the question, or fragments where speakers think aloud), these were replaced with another randomly selected answer. Moreover, lists where the prefinal and final word were separated by a conjunction (i.e., lists of the form "A, B, and C") were replaced as well. In addition, for each speaker two filler items were selected of different lengths. Fillers could include other spoken text (such as repetitions or corrections), and as a result the average length of filler items was 11 words. Each stimulus was cut from the interview session in such a way that it started immediately after the interviewer finished asking the current question until 1000 ms after the speaker finished answering (i.e., 1000 ms after the auditory speech signal of the answerer had stopped).

2. Participants

For the reaction time experiment, 30 right-handed native speakers of Dutch participated, seven male and 23 female, between 24 and 62 years old. None of the participants had participated as a speaker in the data collection phase, and none was involved in audio-visual speech research.

3. Procedure

Stimuli were presented to participants in three conditions: one bimodal one, containing audio-visual stimuli (AV), and two unimodal ones, one audio only (AO), and one vision only (VO). In the audio-visual condition, participants saw the stimuli as they were recorded. In the audio-only condition, participants heard the speakers while the visual channel only depicted a static black screen, and in the vision-only condition, participants only saw the speakers but could not hear them. All participants entered all three conditions (within design), but the order in which participants entered these conditions was systematically varied (using a 3×3 Latin square design). Moreover, within a condition, stimuli were always presented in a different random order. In this way, all potential learning effects could be compensated for.

Each condition consisted of two parts: a baseline measurement and the actual end-of-utterance detection. Each part was preceded by a short practice session so participants would be acquainted with the experimental setting and the kind of stimuli in the current condition. The practice session did not contain lexical material which reoccurred in the actual experiment.

The aim of the baseline measurement was to find out how long it took participants on average to respond to comparable stimuli in the three modalities of interest (AV, AO, VO) of varying durations but always completely devoid of finality cues. During the baseline measurement, the participants' task was to press a designated button as soon as the end of the stimulus was reached. Stimuli were constructed to make them comparable to the actual stimuli used in the nonbaseline conditions but without introducing potential finality cues. In the audio-visual modality, the baseline stimuli therefore consisted of a video still (a single frame of some speakers) accompanied by a stationary /m/ (a male voice for male speakers, and a female voice for female speakers), creating the impression of a speaker uttering a prolonged "mmm." In the vision-only baseline measurement, only the video still was displayed, and in the audio-only baseline measurement, only the stationary /m/ was heard. In all three conditions the baseline stimuli are therefore completely static: the face does not move, since it is a still image, and the sound does not change either, since it is stationary. When the end of a baseline stimulus is reached, the sound stops (in the AO condition) and a blank screen appears (in the VO condition); this happens simultaneously in the AV condition. Only then can participants know that the stimulus ended; there is no conceivable cue in the stimulus which could presignal this.

During the actual end-of-utterance detection part, participants were instructed to indicate, as soon as possible, when the speaker finished his or her utterance by pressing a dedicated button. In the experiment, it was crucial that participants pay attention to visual information on the screen. Therefore, they were given an additional monitoring task, where participants had to press another button as soon as they saw a small red dot appearing on the screen. These red dots were added to a limited number of dummy stimuli. Even though the audio-only condition did not include any potentially relevant visual information (only a black screen), participants also had to spot the red dots in this condition to make sure all conditions were alike in this respect. The duration of the red dot appearance was 1/25 s (a single frame); it appeared at varying locations on the screen. The dummy stimuli were only used to control the visual attention of participants and were not used in the reaction time analyses. This use of dots to make sure participants process visual information is a common procedure in audiovisual speech research (e.g., Bertelson *et al.*, 2003).

The experiment was individually performed. Participants were invited into a quiet room, and asked to take a seat behind a computer on which the stimuli would be displayed. There were loudspeakers to the left and right of the screen through which the sound was played. Participants received instructions before each of the three conditions and before they started with the relevant practice session. If everything was clear, the actual experiment started and the experimenter moved out of the visual field of the participant. There was no further interaction between participant and experimenter during the experiment.

4. Data processing

Reaction times (RTs) were always measured in milliseconds from the actual end of utterance (i.e., the moment where the speech signal ended). An RT of 0 thus means that a participant pressed exactly at the end of the utterance (when the auditory speech signal stopped). Notice that in the baseline measurement, the end of the dummy utterance /mmm/ also marked the end of the stimulus. In the actual experiment, stimuli continued for 1000 ms after the speaker

TABLE I. Reaction times in milliseconds for the different conditions: audio-visual (AV); vision-only (VO); audio-only (AO) in both the baseline measurement and the actual experiment, with standard errors and with 95% confidence intervals.

Measurement	Condition	RT	Std. error	95% CI
Baseline	AV	391.7	7.6	(376.1,407.3)
	VO	330.8	5.9	(318.9,342.9)
	AO	380.3	5.5	(368.9,391.7)
Experiment	AV	508.8	38.6	(429.7,587.8)
	VO	668.5	33.3	(600.4,736.7)
	AO	524.6	40.2	(442.4,606.9)

finished speaking (i.e., after the spoken audio signal ended), and the end of utterance thus does not coincide with the end of the stimulus.

Inspection of the measurements revealed that occasionally a negative RT was recorded. This happened 13 times during the baseline measurement (i.e., 1.8% of the baseline data points), and 302 times during the actual experiment (nearly 7% of the experimental data points). In both cases, the negative RTs were evenly distributed over the modality conditions. In the case of the baseline measurement we can be certain that these are errors, since participants had to respond to the "ending" of the baseline stimuli and, as explained above, there were no cues that could possibly presignal the end. Hence these errors were replaced by the mean RT value for that stimulus. It is important to note that this did not significantly alter the results, so the inclusion of the negative RTs in the baseline condition would have led to basically the same results as reported below (given the very small number of negative instances).

In the actual end-of-utterance experiment a negative RT is not necessarily an error, because here, as noted in Sec. I, presignals may occur, and hence the participant may believe the end of the utterance is near even though the speaker has not actually stopped speaking yet. Since there is no other criterion for their exclusion, we decided not to remove these negative RTs. Finally, there was a total of 23 nonresponses (0.5%), which were treated as missing values in the statistical analysis. We did not manipulate the raw data in any other way.

5. Statistical analyses

All tests for significance were performed with a repeated measures analysis of variance (ANOVA). Mauchly's test for sphericity was used, and when it was significant or could not be determined, we applied the Greenhouse–Geisser correction on the degrees of freedom. For the sake of transparency, we report on the normal degrees of freedom in these cases. *Post hoc* analyses were performed with the Bonferroni method.

B. Results

A general overview of the RT results for the different conditions can be found in Table I. First consider the baseline measurement. Here the VO condition evoked the fastest reaction times followed by the AO and the AV conditions. An ANOVA was performed with condition and stimulus duration as within participants variables and reaction time as the dependent variable was performed. It indeed revealed a main effect of condition $[F(2,58)=11.215, p<0.001, \eta_p^2=0.279]$. *Post hoc* analyses showed that there was a significant difference between the audio-visual and vision-only condition ($p<0.001$), and between the vision-only and the audio-only condition ($p<0.001$). The audio-only and the audio-visual condition did not differ significantly ($p=0.368$). The stimuli used for the baseline measurement differed in duration, but this did not have a significant influence on the reaction times [$F(7,203)=2.891$, *n.s.*], nor was the interaction between condition and stimulus duration significant [$F(14,406)=2.021$, *n.s.*].

TABLE II. Reaction times in milliseconds for the different conditions: audio-visual (AV); vision-only (VO); audio-only (AO) in the actual experiment as a function of length (three words or five words), with standard errors between brackets.

	Length	
Condition	Three words	Five words
AV	585.0 (36.6)	432.5 (42.7)
VO	803.9 (33.0)	533.0 (44.3)
AO	627.6 (48.9)	421.7 (42.6)

Next consider the results of the actual experiment. Here the AV condition yielded the quickest responses, followed by the AO condition, while the VO condition leads to the slowest reaction times. An ANOVA with condition, length (measured by the number of words: three or five), and speaker as within participants variables and reaction time as the dependent variable was carried out. A significant main effect of condition was found $[F(2,58)=17.052, p<0.001, \eta_p^2=0.370]$. *Post hoc* analyses showed that there was a significant difference between the audio-visual and vision-only condition ($p<0.001$), and between the vision-only and the audio-only condition ($p<0.001$). The audio-only and the audio-visual condition did not differ significantly ($p=0.396$). In addition, a main effect of stimulus length was found $[F(1,29)=90.086, p<0.001, \eta_p^2=0.756]$. Inspection of Table II reveals that three word utterances led to longer reaction times than five word utterances. Finally, there was also a main effect of speaker $[F(7,203)=23.500, p<0.001, \eta_p^2=0.448]$ which indicates that some speakers gave overall better or more cues that they were nearing the end of the utterance than other speakers did.

When looking at the interaction effects, a significant interaction between condition and stimulus length $[F(2,58)=26.480, p<0.001, \eta_p^2=0.477]$ was found. As can be seen in Table II, the RT for three word utterances and for five word utterances differs substantially across the different conditions: it is relatively small for the audio-visual condition and relatively large for the vision-only condition, suggesting that the presence of extra cues in longer fragments is particularly useful for the vision-only condition. The RT patterns for the eight speakers are similar over the three modality conditions, as can be seen in Fig. 2. However, some speakers score par-

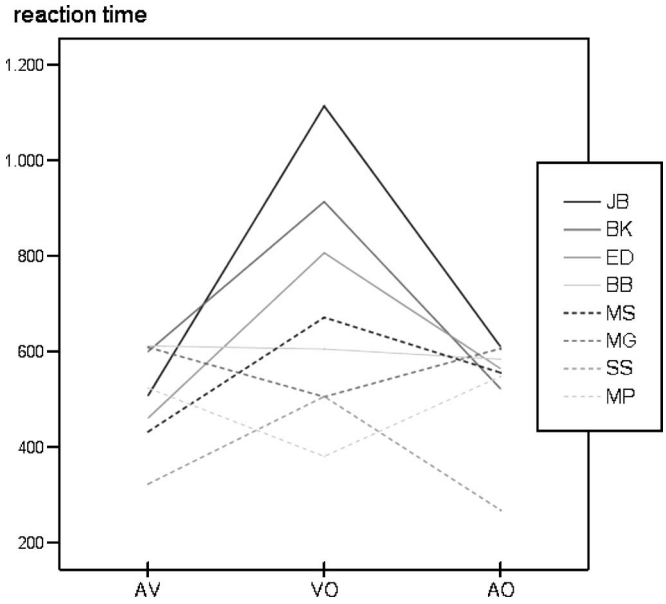

FIG. 2. The mean reaction time (in ms) for the different speakers in the three modalities.

ticularly well in one of the conditions, for instance, because they better cue the end of their utterances using facial cues rather than auditory ones.

It is interesting to see that the reaction time patterns for the baseline measurement are rather different from those of the actual experiment. The aim of the baseline measurement was to find out how long it takes to respond to a stimulus without any finality cues presented in a certain modality, and to compare these scores with the reaction times in the actual experiment in order to eliminate the influence of the presentation modality itself. The picture that emerges is visualized in Fig. 3, which shows that the reaction times for the baseline and nonbaseline versions are more similar in the audio-visual condition, and more divergent in the vision-only condition,

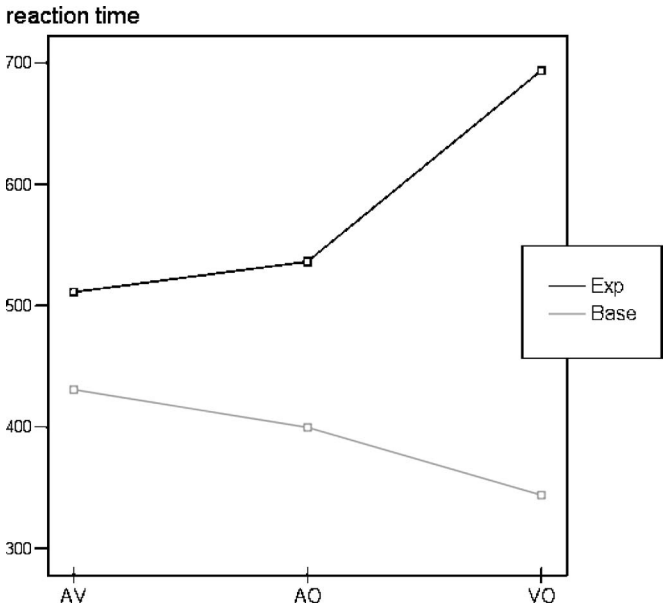

FIG. 3. The mean reaction time (ms) in the three conditions for the baseline and the actual experiment.

while the results for the audio-only condition are inbetween these two extremes. That is, where the visual modality leads to the fastest RT results in the baseline measurement, they are the slowest in the actual experiment. The reverse is true for the data in the audio-visual modality, whereas the data for the auditory modality are in the middle in both sessions.

To test these differences, we computed a difference score for each participant and stimulus, by subtracting the audio-visual baseline RT scores from that participant from his or her nonbaseline RT scores for the audio-visual stimuli, and similarly for the other two modalities. The resulting average difference score was 80.3 ms for the audio-visual condition, 136.8 ms for the audio-only condition, and 349.9 ms for the vision-only condition. We then performed a univariate ANOVA with average difference score for each participant as the dependent variable, and condition (AV, AO, VO) as the independent variable, which indeed revealed a significant effect of condition on difference score $[F(2,87)=13.704, p<0.001, \eta_p^2=0.40]$. A Bonferroni *post hoc* analysis revealed that all pairwise comparisons were significant at the $p<0.001$ level, except for the one between the audio-visual and the audio-only condition ($p=0.906$).

C. Summary

In the first experiment, we measured reaction times for end-of-utterance detection in three different conditions: audio only, vision only, and audio-visual. If prediction of the end of a turn was impossible, the reaction times for the different modalities in the actual experiment would have been the same, or at least have the same pattern as in the baseline measurement, where no cues were present. However, this is clearly not what was found. Rather, the audio-visual stimuli in the actual experiment led to the quickest responses, the audio-only stimuli led to slightly longer reaction times (although the difference with the audio-visual stimuli was not statistically significant), and the vision-only stimuli led to the slowest responses. While this result suggests that combining modalities is useful for end-of-utterance detection, it also leaves open the possibility that participants essentially rely on auditory information only for end-of-utterance detection. This issue is investigated more closely in a second experiment, where participants have to classify brief fragments as nonfinal or final (end of utterance) ones.

IV. EXPERIMENT II: CLASSIFICATION

The design of the classification task experiment resembles the design used in gating tasks. In a gating task a spoken language stimulus is presented in segments of increasing duration, usually starting at the beginning of the stimulus. Participants must try to recognize the entire spoken stimulus on the basis of the fragment (Grosjean, 1996).

In one possible presentation format, the *duration-blocked format*, participants are presented with all the stimuli at a particular segment size, then all the stimuli again in a different segment size (Grosjean, 1996; Walley *et al.*, 1995). In the current experiment we used two sizes, a long and a short one, both of which did not cover the entire original utterance. Participants had to make a binary decision about

the setting from which the fragment originated (i.e., final or not final).

A. Method

1. Stimuli

The stimuli for Experiment II were selected from the utterances of the same eight speakers which were used in Experiment I. For each of these speakers we randomly extracted answers from their original set of answers (see Sec. II), and constructed two types of fragments from these: short ones, consisting of one word, and long ones, consisting of two words. Orthogonal to this, half of the fragments were from a final (end-of-utterance) and half from a nonfinal position. In the same way as for Experiment I, we made sure that participants could not pick up on lexical cues for their final/nonfinal classifications.

For each of the eight speakers, we created four short pairs (final/nonfinal) and four long pairs of fragments, where the short fragments always consisted of the last word of the corresponding long (two word) fragment. Naturally, the final pairs were always selected from the tail of the list, while the nonfinal pairs were selected from varying positions in the list. The length of the original context surrounding a fragment was more or less balanced, with a small majority of fragments extracted from answers consisting of five words.

To guarantee the understandability of the fragments and to make sure they were comparable across conditions, the fragments were selected such that they included a naturally occurring pause after the last word of the fragment (when it was a nonfinal fragment), or a pause after the end of the original answer (when it consisted of the final part of an answer). The fragments were always cut in such a way that the pauses in the corresponding one word and two word stimuli lasted equally long, to make sure that the length of the pause (which, as noted in Sec. I, is an important signal for end of utterance) could not be used as a cue for classification.

As for Experiment I, all fragments were stored in three ways: AO, VO, or AV. Therefore, in total 128 stimuli were created for each modality: 8 speakers $\times 2$ lengths (short-long) $\times 2$ types (nonfinal and final) $\times 4$ fragments.

2. Participants

The participants consisted of a group of 60 native speakers of Dutch; 25 male and 35 female, between 20 and 56 years old. None of them participated as a speaker in the data collection phase nor as a participant in Experiment I, and none was involved in audio-visual speech research.

3. Procedure

Participants were given a simple classification task: they were told to determine for each fragment whether it marked the end of a speaker's utterance or not. Again, stimuli were presented in three conditions: an AV, an AO, and a VO, which were presented to participants in the same format as in Experiment I, but this time in a between-participants design.

Each condition consisted of two parts: one part for the short (one word) fragments and one part for the long (two word) fragments. The order in which participants passed the two different parts was systematically varied. For each part, two lists were created with a different random order. Participants were exposed to either the A versions or the B versions of a list. Therefore, each participant passed the items in a different random order in each part, and since the order in which participants underwent the short and long fragments part was also systematically varied, potential learning effects could be compensated for.

Each condition was preceded by a short practice session, consisting of two stimuli (different from the experimental stimuli), so that participants could get used to the type of tasks and stimuli. The general procedure was the same as for Experiment I.

TABLE III. For each factor, the levels of the factor, the percentage of correct judged utterances with standard errors, and 95% confidence intervals are given.

Factor	Level	Percent correct (%)	Std. error	95% CI
Fragment type	NF	80.8	0.11	(78.6,83.0)
	F	75.2	0.12	(72.9,77.7)
Stimulus length	Short	75.1	0.09	(73.3,77.0)
	Long	81.0	0.07	(79.5,82.3)
Modality	AV	84.7	0.11	(82.5,86.9)
	VO	75.7	0.11	(73.6,77.9)
	AO	73.6	0.11	(71.5,75.8)

4. Statistical analyses

Tests for significance were performed with a repeated measures ANOVA with speaker (eight levels), stimulus length (short: one word, long: two words), and fragment type (not final, final) as within-subjects factors and modality (VO, AO, AV) as a between-subjects factor (mixed design) and with the percentage of correct classifications over the four fragments as the dependent variable (recall that for each speaker four short and long pairs of final and nonfinal stimuli were selected). Mauchly's test for sphericity was used to test for homogeneity of variance, and when this test was significant or could not be computed, we applied the Greenhouse–Geisser correction on the degrees of freedom. For the purpose of readability, we report the normal degrees of freedom in these cases. The Bonferroni correction was applied for multiple pairwise *post hoc* comparisons, and contrasts were computed in several cases.

B. Results

Table III gives the overall results for three factors of interest, i.e., fragment type, stimulus length, and modality. According to the ANOVA all three factors had a significant influence on the classification. First, consider the main effect of fragment type $[F(1,57)=7.855, p<0.01, \eta_p^2=0.121]$. It appears that judging nonfinality is somewhat easier than judging finality (80.8% versus 75.2%), but overall it is clear that the vast majority of fragments are classified correctly.

Stimulus length also had a significant influence $[F(1,57)=28.800, p<0.001, \eta_p^2=0.336]$. Inspection of Table

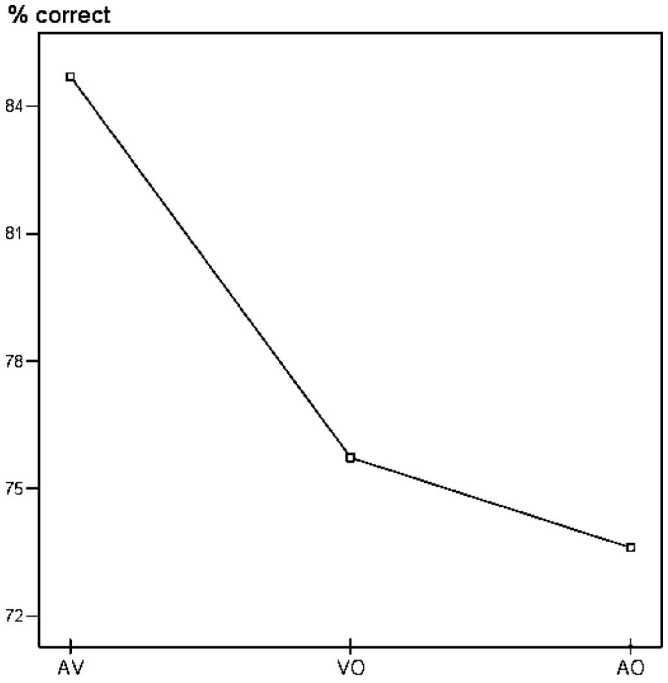

FIG. 4. Percentage of correct answers in the audio-visual (AV), vision-only (VO), and audio-only (AO) conditions.

III reveals that short (one word) fragments are somewhat more difficult than longer (two word) fragments.

The most interesting main effect is that of modality, which was significant as well [$F(2,57)=29.475, p<0.001, \eta_p^2=0.508$]. It is interesting to note that both unimodal conditions yield around 75% correct classifications (75.7 for the vision-only condition and 73.6 for the audio-only condition), and that both are clearly outperformed by the bimodal, audio-visual condition (with 84.7% correct). *Post hoc* analyses showed that there was a significant difference between the audio-visual and the vision-only condition ($p<0.001$), and between the audio-visual and the audio-only condition ($p<0.001$). The vision-only and the audio-only condition did not, however, differ significantly ($p=0.54$). This pattern of results is visualized in Fig. 4.

Besides the main effects for the three factors listed in Table III, the factor speaker also had a significant main effect [$F(7,399)=52.375, p<0.001, \eta_p^2=0.48$]. As can be seen in Table IV, the total number of correct classifications differs per speaker, ranging from 63% correct for speaker JB to 87.8% for speaker SS. *Post hoc* analyses showed that this difference was significant ($p<0.001$). Various other pairwise comparisons of speakers were significant as well, and this shows that there are overall substantial differences between speakers in end-of-utterance signaling. It is rather interesting to observe that the scores per speaker may differ across conditions. Indeed, a significant two-way interaction was found between speaker and modality [$F(7,399)=14.764, p<0.001, \eta_p^2=0.341$]; in Table IV it can be seen that, for instance, speaker BB apparently offers clearer visual than auditory cues, as the percentage of correctly classified stimuli for this speaker drops considerably in the AO condition. This is different for speaker MG, for instance, who seems to send more useful auditory cues (in her case the classification scores drop in the VO condition). Simple contrasts showed that this difference was significant [$F(2,57)=78.839, p<0.001, \eta_p^2=0.734$].

TABLE IV. For each speaker, the total percentage of correctly judged utterances, and the percentage of correctly judged utterances as a function of the three modalities.

Speaker	AV	VO	AO	Total
BB	86.5	86.5	56.8	76.7
BK	74.1	74.4	59.3	69.3
ED	90.6	73.3	77.7	80.5
JB	64.7	57.5	66.9	63.0
MG	86.6	68.1	86.0	80.2
MP	85.9	76.7	76.2	79.6
MS	93.1	87.2	81.0	87.1
SS	96.2	82.0	85.0	87.8

TABLE V. For each modality, the percentage of correctly judged utterances, as a function of stimulus length (one or two words) and fragment type (nonfinal and final).

Length	Finality	AV	VO	AO	Total
1	NF	81.8	76.2	69.7	75.9
1	F	83.1	73.6	66.0	74.3
Subtotal		82.5	74.9	67.9	
2	NF	89.4	82.6	85.2	85.7
2	F	84.5	70.6	73.6	76.2
Subtotal		86.9	76.6	79.4	
Total		84.7	75.7	73.6	

In addition, a significant two-way interaction was found between fragment type and stimulus length [$F(1,57)=11.317, p<0.01, \eta_p^2=0.166$]. This interaction can also be explained by looking at Table V, where it can be seen that for the nonfinal fragments, the longer stimuli evoked more correct answers (85.7%) than the short stimuli (75.9%), while for the final fragments the stimulus length makes almost no difference (74.3% versus 76.2%, respectively).

Table V also illustrates a second significant two-way interaction between stimulus length and modality [$F(2,57)=6.889, p<0.01, \eta_p^2=0.195$]. As expected, for both stimulus lengths, the audio-visual modality is the easiest one. For the short fragments, the audio-visual modality (82.5% correct answers) is followed by the visual modality (74.9%), and subsequently the auditory modality (67.9%). A *post hoc* test within the short word fragments revealed that all pairwise comparisons are statistically significant (AV-VO, $p<0.01$, AV-AO $p<0.001$, and VO-AO, $p<0.05$). However, for the long fragments, the audio-visual modality (86.9% correct answers) is followed by the auditory modality (79.4%), and subsequently the visual modality (76.6%). A *post hoc* within the long fragments revealed that all pairwise comparisons differ at the $p<0.001$ level, with the exception of the difference between VO and AO which is not significant. No other significant interactions were found.

C. Summary

The classification experiment reveals that speakers can make the best end-of-utterance classifications for bimodal,

audio-visual stimuli. It is interesting to observe that the numerically lowest scores are obtained for the audio-only condition, which has received the most attention in the literature. The vision-only results are somewhat better, which shows that visual cues to end of utterance are indeed useful for participants. Besides the modality effects, some other interesting results were obtained. A small response bias was found for nonfinal fragments, so that nonfinal fragments are slightly more often classified correctly. For the nonfinal fragments, the longer stimuli evoked more correct answers than the short stimuli, while for the final fragments the stimulus length makes almost no difference. Finally, the classification scores were found to vary per speaker, both overall and as a function of modality.

V. GENERAL DISCUSSION AND CONCLUSION

The fact that speakers use auditory cues (intonation, pausing, rhythm, etc.) which indicate that they are approaching the end of their utterance is well established (e.g., de Pijper and Sanderman, 1994; Price *et al.*, 1991; Swerts *et al.*, 1994a; 1994b; Wightman *et al.*, 1992). Various researchers have pointed out that speakers may also employ visual cues (such as posture, head movements, or gaze) for this purpose (e.g., Argyle and Cook, 1976; Cassell *et al.*, 2001; Nakano *et al.*, 2003; Vertegaal *et al.*, 2000). While the auditory cues have been studied from a perceptual perspective as well, comparable studies addressing the perception of visual cues (or the audio-visual combination) for end-of-utterance detection are thin on the ground. This naturally raises the question which modalities people actually employ to determine whether a speaker is at the end of an utterance and what the effect is of combining information from different modalities. In order to answer these questions, we first collected utterances in a semispontaneous way using a new experimental paradigm eliciting target list answers of three or five words long, making sure that target words could occur at the beginning, middle, or end of the list. On the basis of these utterances, two perception experiments were carried out.

As a first exploration, we performed a reaction time experiment in which participants were confronted with utterances, taken out of their original interview context to make sure that participants could not rely on lexical cues, and presented in three formats: VO, AO, or AV. The task for participants was to indicate as soon as possible when the speaker reached the end of his or her current utterance. It was found that participants could do this most quickly in the bimodal, audio-visual condition, followed (with a relative small, nonsignificant margin) by the audio-only condition, and with the slowest responses in the vision-only condition.

To find out how participants respond to stimuli in the respective conditions without any cues that participants might relate to (non)finality, we also performed a baseline reaction time measurement using artificially created static stimuli. Even though these artificial stimuli are of necessity not fully comparable with the real, experimental stimuli, comparing the experimental scores with those obtained in the baseline reveals some suggestive differences. It is interesting to observe that in the baseline condition, the audio-visual stimuli led to the slowest responses. That RTs for the AV condition are slower in the baseline than in the actual experiment may be explained by the thesis that when two different modalities (which contain no cues when their presentation will end) are offered at the same time, they will produce a cognitive overload because two sources of information have to be processed instead of one (Doherty-Sneddon *et al.*, 2001). However, when two modalities are presented in a situation where the information does contain predictive cues, as in the nonbaseline condition, the different modalities might serve as sources providing complementary information, and thus can help each other in resolving ambiguous slots in the stream of speech (compare Kim *et al.*, 2004; Schwartz *et al.*, 2004).

In general, the responses to the baseline stimuli were substantially faster than the responses in the nonbaseline conditions. This is in line with various reaction time studies concluding that a complex stimulus leads to slower reaction times (e.g., Brebner and Welford, 1980; Luce, 1986; Teichner and Krebs, 1974). Since the baseline stimuli are essentially static, without any variations that might be informative for end-of-utterance detections, there is much less information to process than in the experimental stimuli.

It was also interesting to see that the five word stimuli lead to quicker responses than the three word ones, which is in line with the studies of Carlson *et al.* (2005) and Swerts and Geluykens (1994). Again, this result is also consistent with findings from the literature on reaction time studies. Froeberg (1907), for instance, already found that longer visual stimuli elicit faster reaction times than stimuli of a shorter duration, and Wells (1913) found the same for auditory stimuli. In general, it is known that stimulus duration has a clear impact on reaction times (e.g., Ulrich *et al.*, 1998). Moreover, in this particular setup, the five word stimuli may also simply contain more potential finality cues than the three word stimuli, which would be an additional explanation for the fact that five word stimuli result in quicker responses than three word ones.

The results from the first experiment cannot be used to rule out the possibility that auditory information is sufficient for end-of-utterance detection, since it did not result in a significant difference between the audio-visual and the audio-only condition. Therefore a second experiment was conducted, to get more insight into how participants respond to stimuli in the different modalities. In this experiment participants were offered short (one word) and long (two word) fragments which either did or did not mark the end of an utterance, and participants had to classify these as final or nonfinal. In this experiment the bimodal presentation format gave significantly better results than the unimodal ones: when participants have access to both auditory and visual cues they make more adequate classifications than in situations where they only have information from one modality at their disposal. It was interesting to observe that overall most mistakes were made in the audio-only condition, i.e., the situation which has received the most attention in the literature so far, although the difference between the respective unimodal conditions was not statistically significant. Two possible explanations can be given for the superiority of the

audio-visual stimuli in this particular experiment. First, a combined audio-visual presentation format clearly offers more cues than a presentation in a single modality. But we have also seen that speakers differ in which signals they give, with some speakers showing more visual cues and others more auditory ones. Clearly, this also speaks in favor of a bimodal presentation.

In addition a slight response bias was found for nonfinal fragments, with nonfinal fragments more often classified correctly than the final ones. And for the nonfinal fragments, it was found that the longer stimuli were more often classified correctly than the shorter ones, while stimulus length did not have an effect on the final fragments. This suggests that when finality cues are available, it makes no difference whether the fragment is short or long, but when finality cues are not available, participants need longer fragments to make a decision. This could be caused by the fact that finality is displayed in local cues, thus in the last part of a fragment, just before it stops. In contrast, when no local finality cues are displayed, people need to base their decision more on global cues. In general, it is a well-known finding in cognitive psychology that it is easier to determine whether a cue is present than to decide that something is not there (e.g., Hearst, 1991).

It is also noteworthy that the longer fragments are better classified than the short fragments in the audio-only condition, which suggests that the finality cues in speech seem to be more global in nature, and hence that participants can make better judgments for longer fragments when more of these global cues are available. For the vision-only condition, length does not appear to have an influence, which suggests that the visual cues may be more local. Notice that this would also offer an explanation for the fact that the audio-only condition outperforms the vision-only condition in Experiment I, but not in Experiment II. Since the stimuli in the second experiment were overall shorter (consisting of one or two words) than those in the first experiment (which consisted of entire utterances of three or more words), the participants in the second experiment could not use the spoken global cues to the full effect.

The focus in this paper has been on a perceptual comparison of the cue value of different modalities for signaling end of utterance. However, it would be interesting to see which auditory and visual behaviors might have served as cues in both experiments. To gain some insight into this, we annotated for both the final and the nonfinal stimuli the 50% that received the best classification scores in Experiment II. In particular, we concentrated on those cues that are known from the literature (see Sec. I), and that could clearly and consistently be determined on the basis of visual or auditory inspection of our stimuli. The following auditory cues were labeled:

(1) **Boundary tone:** whether a fragment ends in a low (*L*), medium (*M*), or high boundary tone (*H*); and

TABLE VI. Representative stills illustrating the annotated visual features. Notice that various stills contain multiple features, since cues may cooccur. For example, the female speaker with her mouth open also moves her head and eyes away.

Label	Example	
Brows [up]		
Eyes [away]		
Mouth [open]		
Head [away]		
Posture [away]		

(2) **Creaky voice:** whether a stimulus contains some creaky fragments.

In both cases, the annotation was determined by perceptual judgments, and performed by professional intonologists. The distinction between high, mid, and low boundary tones was determined by comparing the tonal pattern in the final syllables of the fragment to the pitch range of the preceding part. If the final stretch of speech was clearly below or above the preceding pitch range, it would be categorized as either low or high, whereas a tone inbetween those two extremes would get a mid label.

In the visual domain, the following features were labeled (Table VI contains representative stills for each of the visual features):

TABLE VII. The annotation as a function of fragment type (nonfinal and final).

Modality	Feature	Setting	NF	*F*	Total
Auditory	Boundary tone	*H*	0	6	6
		M	13	2	15
		L	3	8	11
	Creaky voice		5	5	10
Visual	Brows	Up	11	8	19
		Down	3	4	7
	Eyes	Blinking	7	12	19
		Away	23	8	31
		Back	3	13	16
	Mouth	Open	6	2	8
		Closed	0	4	4
	Head	Nodding	12	21	33
		Away	10	4	14
		Back	1	4	5
	Posture	Away	7	6	13
		Back	0	2	2

(1) **Brows:** whether the eyebrows are raised (up) or lowered (down);
(2) **Eyes:** whether the eyes of the speaker are turned away from the camera (away), or whether the speaker returns his/her gaze towards the camera (back); we also labeled cases where a speaker was blinking;
(3) **Mouth:** whether the mouth at the end of the stimulus is closed or open;
(4) **Head:** whether the speaker turns his/her head away from the camera during the answer, or moves the head back to the camera; moreover, we also labeled cases where the speaker makes a nodding movement during the fragment; and
(5) **Posture:** whether the speaker changes his/her posture away from the camera, or rather moves his/her body back towards the camera.

The cues were always labeled blind to condition, in order to avoid circularity in their annotation. Table VII gives the overall results for the factors of interest, split by the two possible modalities, i.e., auditory (boundary tones, creaky voice) and visual (brows, eyes, mouth, head, posture) as a function of fragment type (nonfinal of final).

In the auditory domain, it can be observed that the midending tones are more typical for the nonfinal fragments, while both high and low boundary tones occur more often at the end of final fragments. This result is in line with many previous studies which show that a clearly low or high tone (such as in question intonation) may signal the end of an utterance, whereas a midtone serves to cue continuity (e.g., Caspers 1998; Silverman and Pierrehumbert, 1990). At first sight, the presence of a creaky voice (which in our stimuli rarely happens in the first place) does not appear to be related to finality or nonfinality, but a closer inspection of the stimuli reveals that all the noncreaky fragments occur in cases where speakers used a midtone, while the creaky fragments only occur when speakers produce a high or low tone, so that creakiness may serve as an extra cue to reinforce the finality/nonfinality marking of boundary tones. With respect to the visual features, Table VII suggests that there is a clear tendency for speakers to divert their eyes and head in nonfinal fragments, while they return eyes, head, and also posture in the final fragments. Additionally, there is a trend for the mouth to be still open when a fragment has not yet been finished (even though the speaker is not speaking), whereas a mouth is more often closed at the end of a final fragment. Also, final fragments display relatively more cases of blinking and nodding, while the brows tend to be up or down at the end of nonfinal versus final fragments, respectively.

There are also many individual differences between speakers. In the annotated utterances, speakers produce almost 23 cues on average, but there are clear differences. Speaker JB for instance, produces only 14 visual cues to signal finality, which is consistent with the fact that speaker JB was most difficult to classify in Experiment II. On the other hand, speaker JB tends to use low boundary tones more often than other speakers. This may account for the observation, for Experiment I, that participants took relatively long to respond to JB's stimuli in the vision-only modality, and were rather quick for this speaker in the audio-only and audio-visual conditions. Speaker SS, to give a second example, is visually the most expressive (33 visual cues) and indeed her stimuli lead to the overall quickest responses in Experiment I, and to the most correct classifications in Experiment II. Apart from the fact that some speakers display more cues than others, some speakers also tend to display different cues than other speakers. For example, on the visual level, while most speakers return their gaze in a final position, some speakers (e.g., ED) do not return their gaze but instead nod more often in the final position.

This small scale annotation reveals that many of the cues mentioned in Sec. I indeed occur in the stimuli, and it seems likely that participants made their classification on the basis of these various cues. In future research, it would be interesting to find out how the different audio-visual features discussed above are distributed over the whole utterance. It has been argued (Argyle and Cook, 1976) that an utterance consists of different phases, i.e., a starting phase, a middle phase, and a closing phase, which are connected to patterns in eye gaze (see also Cassell *et al.*, 2001 for similar kinds of observations in other bodily gestures.) It remains to be seen whether such patterns are also true for other visual features, and how these relate to more global auditory cues, such as declination or rhythmic patterns. It would also be interesting to test the relative importance of the various auditory and visual cues in followup experiments.

In sum: our study, using a reaction-time experiment and a classification task, has revealed that subjects are sensitive both to auditory and visual signals when they need to estimate whether or not a speaker utterance has ended. While both modalities separately contain cues that enable subjects to make reliable finality judgments, it turns out that a bimodal, audio-visual condition leads to the most accurate results. The relative cue value of the two unimodal conditions depends on the experiment, where auditory cues were more important in the RT experiment, and visual cues in the classification task. In addition, its relative importance also differs

between stimuli from different speakers, due to the fact that some speakers display more auditory cues, and others more visual ones.

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). We thank Lennard van de Laar for various kinds of technical assistance, Carel van Wijk for statistical advice, and Jean Vroomen for allowing us to make use of the Pamar software. We greatly benefitted from the comments of three anonymous reviewers on a previous version of this paper.

Argyle, M., and Cook, M. (**1976**). *Gaze and Mutual Gaze* (Cambridge University Press, Cambridge, UK).

Beattie, G. W., Cutler, A., and Pearson, M. (**1982**). "Why is Mrs. Thatcher interrupted so often?" Nature (London) **300**, 744–747.

Bertelson, P., Vroomen, J., and de Gelder, B. (**2003**). "Visual recalibration of auditory speech identification: A McGurk aftereffect." Psychol. Sci. **14**, 592–597.

Brebner, J., and Welford, A. (**1980**). "Introduction: An historical background sketch," in *Reaction Times* edited by A. Welford (Academic, New York), pp. 1–23.

Carlson, R., Hirschberg, J., and Swerts, M. (**2005**). "Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates." Speech Commun. **46**, 326–333.

Caspers, J. (**1998**). "Who's next? The melodic marking of questions vs. continuation in Dutch." Lang Speech **41**, 375–398.

Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., and Rich, C. (**2001**). "Non-verbal cues for discourse structure." *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL)*, Toulouse, France, July 9–11, pp. 114–123.

Couper-Kuhlen, E. (**1993**). *English Speech Rhythm* (Benjamins, Philadelphia).

de Pijper, J. R., and Sanderman, A. A. (**1994**). "On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues." J. Acoust. Soc. Am. **96**, 2037–2047.

de Ruiter, J. P., Mitterer, H., and Enfield, N. (**2006**). "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation." Language **82**, 515–535.

Doherty-Sneddon, G., Bonner, L., and Bruce, V. (**2001**). "Cognitive demands of face monitoring: Evidence for visuospatial overload." Mem. Cognit. **29**, 909–917.

Doughty, M. J. (**2001**). "Consideration of three types of spontaneous eyeblink activity in normal humans: During reading and video display terminal use, in primary gaze, and while in conversation." Optom. Vision Sci. **78**, 712–725.

Duncan, S. (**1972**). "Some signals and rules for taking speaking turns in conversations." J. Pers Soc. Psychol. **23**, 283–292.

Ekman, P. (**1979**). "About brows: Emotional and conversational signals," in *Human Ethology: Claims and Limits of a New Discipline*, edited by M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog (Cambridge University Press, Cambridge, UK), pp. 169–202.

Froeberg, S. (**1907**). "The relation between the magnitude of stimulus and the time of reaction." Arch. Psychol. (Frankf) **8**.

Goodwin, C. (**1980**). "Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning." Sociological Inquiry **50**, 272–302.

Grosjean, F. (**1983**). "How long is the sentence? Prediction and prosody in the on-line processing of language." Linguistics **21**, 501–529.

Grosjean, F. (**1996**). "Gating." Lang. Cognit. Processes **11**, 597–604.

Hearst, E. (**1991**). "Psychology and nothing." Am. Psychol. **79**, 432–443.

Kendon, A. (**1967**). "Some functions of gaze-direction in social interaction." Acta Psychol. **26**, 22–63.

Kim, J., Davis, C., and Krins, P. (**2004**). "Amodal processing of visual speech as revealed by priming." Cognition **93**, B39–B47.

Kobayashi, H., and Kohshima, S. (**1997**). "Unique morphology of the human eye." Nature (London) **387**, 767–768.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y. (**1998**). "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs." Lang Speech **41**, 295–321.

Krahmer, E., and Swerts, M. (**2004**). "More about brows," in *From Brows to Trust: Evaluating Embodied Conversational Agents*, edited by C. Pelachaud and Zs. Ruttkay (Kluwer, Dordrecht), pp. 191–216.

Leroy, L. (**1984**). "The psychological reality of fundamental frequency declination." *Antwerp Papers in Linguistics* (Antwerp University Press, Antwerp, Belgium), Vol. **40**.

Levinson, S. (**1983**). *Pragmatics* (Cambridge University Press, Cambridge, UK).

Luce, R. (**1986**). *Response Times: Their Role in Inferring Elementary Mental Organization* (Oxford University Press, New York).

Maynard, S. K. (**1987**). "Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation." J. Pragmat. **11**, 589–606.

Nakano, Y. I., Reinstein, G., Stocky, T., and Cassell, J. (**2003**). "Towards a model of face-to-face grounding." *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, July 7–12, pp. 553–561.

Novick, D. G., Hansen, B., and Ward, K. (**1996**). "Coordinating turn-taking with gaze." *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, PA, October 3–6, pp. 1888–1891.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, S. (**1991**). "The use of prosody in syntactic disambiguation." J. Acoust. Soc. Am. **90**, 2956–2970.

Schwartz, J.-L., Berthommier, F., and Savariaux, C. (**2004**). "Seeing to hear better: Evidence for early audio-visual interactions in speech identification." Cognition **93**, B69–B78.

Silverman, S., and Pierrehumbert, J. (**1990**). "The timing of prenuclear high accents in English," *Laboratory Phonology: Between the Grammar and Physics of Speech*, edited by J. Kingston and M. Beckman (Cambridge University Press, Cambridge, UK), Vol. **I**, pp. 71–106.

Swerts, M. (**1997**). "Prosodic features at discourse boundaries of different strength." J. Acoust. Soc. Am. **101**, 514–521.

Swerts, M. (**1998**). "Filled pauses as markers of discourse structure." J. Pragmat. **30**, 485–496.

Swerts, M., Bouwhuis, D., and Collier, R. (**1994a**). "Melodic cues to the perceived finality of utterances." J. Acoust. Soc. Am. **96**, 2064–2075.

Swerts, M., Collier, R., and Terken, J. (**1994b**). "Prosodic predictors of discourse finality in spontaneous monologues." Speech Commun. **15**, 79–90.

Swerts, M., and Geluykens, R. (**1994**). "Prosody as a marker of information flow in spoken discourse." Lang Speech **37**, 21–43.

Teichner, W., and Krebs, M. (**1974**). "Laws of visual choice reaction time." Psychol. Rev. **81**, 75–98.

Ulrich, R., Rinkenauer, G., and Miller, J. (**1998**). "Effects of stimulus duration and intensity on simple reaction time and response force." J. Exp. Psychol. Hum. Percept. Perform. **24**, 915–928.

Vertegaal, R., Slagter, R., van de Veer, G., and Nijholt, A. (**2000**). "Why conversational agents should catch the eye." *Proceedings of the Internation Computer-Human Interaction Conference (CHI)*, The Hague, The Netherlands, April 1–6, pp. 257–258.

Walley, A. C., Michela, V., and Wood, D. (**1995**). "The gating paradigm: Effects of presentation format on spoken word recognition by children and adults." Percept. Psychophys. **57**, 343–351.

Ward, N., and Tsukahara, W. (**2000**). "Prosodic features which cue back-channel responses in English and Japanese." J. Pragmat. **23**, 1177–1207.

Wells, G. (**1913**). "The influence of stimulus duration on RT." Psychol. Monogr. **15**.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. (**1992**). "Segmental durations in the vicinity of prosodic phrase boundaries." J. Acoust. Soc. Am. **91**, 1707–1717.