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

Pashiera Barkhuysen, Emiel Krahmer, Marc Swerts*

Communication & Cognition, Tilburg University†

(Dated: May 23, 2006)

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. In this paper we report on a series of experiments aimed at finding out to what extent humans rely on auditory and/or on visual cues for end-of-utterance detection. We first collected speaker utterances via a novel semi-controlled production experiment. The data thus collected were used in two perception experiments, where we 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 quickest in end-of-utterance detection when confronted with bimodal stimuli, followed by respectively the audio-only and vision-only stimuli. Experiment II is a classification experiment, and showed that participants make the most adequate end-of-utterance classifications in, again, the bimodal condition. It is interesting to observe that the lowest scores for this task are obtained in the audio-only condition.

PACS numbers: 42.66.Si, *43.71.-k, 43.71.+m, *43.70.-h, 43.70.+i
Keywords: Audiovisual speech, prosody, end-of-utterance detection, speech production, speech perception

I. INTRODUCTION

Speakers use non-lexical features to demarcate various kinds of speech units, 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. In addition to features that are encoded in the speech signal itself, there is also work into how particular visually observable variation from a speaker's face, like gaze patterns, or other bodily gestures can be used as boundary cues. However, not much is known about the relative cue value of auditory and visual cues for demarcative purposes. Therefore, the aim of this article is to get more insight into which modalities speakers use to indicate the end of their utterance, and how sensitive observers are to these signals. In particular, our goal is to investigate the relative contribution of visual signals compared to auditory signals, whereby we specifically focus on end-of-utterance marking. Before we embark upon the more specific research questions and the experimental set-up of our different studies, we will first review some of the relevant previous work on auditory and visual boundary markers.

There are various auditory cues which have been shown to serve as boundary markers of speech utterances. One of the strongest prosodic indicators of a break is a pause, either a silent interval or a filler such as "uh" and "uhm," as shown by, among others, De Pijper & 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 in between two consecutive speaker turns. Turn switches generally proceed remarkedly smooth, often without any overlap between speakers and with a minimal delay (Koiso et al. (1998); Levinson (1983); Ward & Tsukuhara (2000)).

One of the reasons why the turn-taking mechanism may proceed so fluently, is that speakers "pre-signal" the end of their utterances. Listeners may pick up these cues and therefore may know in time when the current turn will be finished. Various researchers have looked in detail in the nature of these cues. It has been suggested, for instance, that the capacity of listeners to feel an upcoming boundary is based on what is called rhythmic expectancy, which would steer turn-taking to some extent (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 (Price et al. (1991); Wightman et al. (1992)). 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, where it has been shown that tones which reach a speaker's bottom range clearly function as finality cues (Caspers (1998); Koiso et al. (1998); Swerts et al. (1994)). In addition, there appear to exist melodic structuring devices

*Communication & Cognition, Faculty of Arts, Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands; Electronic address: {P.N.Barkhuysen;E.J.Krahmer;M.G.J.Swerts}@uvt.nl
†URL: http://foap.uvt.nl/

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 has been claimed to 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 & 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.

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. (1994) and Swerts & Geluykens (1994) reported that people are able, on the basis of melodic cues, to judge the 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 materials, even when they had to base such estimations on stimuli which consisted of only one word.

It seems safe to conclude that speakers produce and listeners are sensitive to auditory cues marking the end of an utterance. Various researchers have argued that speakers may use visual cues for this purpose as well, where most studies have investigated how various bodily gestures are used as markers of discourse boundaries. First, different studies focused on general changes in posture (Argyle & Cook (1976); Beattie et al. (1982); Cassell et al. (2001); Duncan (1972)). Reportedly, 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 quite some scholarly attention is related to movements of the eyes. Argyle & 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 analysing 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 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 surround (sclera) to the dark iris. It has been argued that this contrast may have evolved to enhance gaze processing (Kobayashi & 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 which may be important for demarcation purposes as well, such as head nods (e.g., Maynard (1987)), eyebrow movements (e.g., Ekman (1979); Krahmer & 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 verbal and visual features. However, there are still a large number of unsolved questions regarding the cue value of such features. There is almost no research which explores how important visual cues are when compared to auditory cues for marking the end of a speech unit. While it has been shown that listeners are very accurate in judging the end of a unit based on speech-only stimuli, we do not know whether they would be equally capable to do so on the basis of only visual features as well. And if so, it is still an empirical question as to how possible visual boundary markers relate to the auditory ones, which of the two have stronger cue value, 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 which are both based on perceptual judgments of stimuli in one of three conditions: vision-only, audio-only or in a bimodal (audio-visual) condition. The 2 experiments make use of audio-visual recordings of semi-spontaneous utterances that were naturally elicited in a question-answering paradigm. The first experiment is 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, has ended. The second experiment makes use of basically the same stimuli as the ones from the first experiment, but this time participants simply have to decide in a binary fashion whether or not a turn has ended; 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 cues to finality.

The rest of this paper is organized as follows. In section II, we describe the procedure to obtain the audio-visual recordings, which we used as a basis for creating the stimuli of our perception experiments. Then we will discuss the two experiments in section III and IV respectively, including sections on procedure, results and discussion of specific findings. We end this paper in section V with a more general reflection on the significance of the results for our understanding of audio-visual boundary marking.

II. AUDIO-VISUAL RECORDINGS

We gathered digital video recordings of speakers responding to questions in a natural, interview-style situation. The questions were intended to evoke lists of words, for instance based on general knowledge (e.g., Q: What are the colors of the Dutch flag? A: Red, white, blue.) or a a set of numbers (e.g., Q: What are the odd numbers between three and fifteen in reversed order? A: Thirteen, eleven, nine, seven, five.) The correct target answers varied in length, consisting of sequences of 3 or 5 words. The interview consisted of 33 questions, of which 25 were experimental and 8 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?).

A total of 22 speakers participated (13 male and 9 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 was collected. The original recordings were made with a digital video camera (25 frames per second). They were subsequently read into a computer and orthographically transcribed.

III. EXPERIMENT I: REACTION TIMES

In this section, we report on a reaction time experiment with the intention to determine the relative contribution of visual and auditory cues, alone and in combination, for end-of-utterance detection.

A. Method

1. Stimuli

For this experiment 4 male and 4 female speakers were randomly selected from the corpus of 22 speakers described above. Per speaker, 3 instances of answers consisting of 3 words and 3 instances of 5 words were randomly selected on the basis of the transcriptions. In addition, for each speaker 2 filler items were selected of different lengths. Fillers could also include other spoken text (such as repetitions of the question or fragments where speakers think aloud). 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 1000ms 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, 7 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

The experiment had a counterbalanced within-participants design, consisting of 3 conditions: one bimodal, 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 speaker 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, but the order in which participants entered these conditions was systematically varied. Within a condition, stimuli were always presented in a different random order. 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 to make participants acquainted with the experimental setting and the kind of stimuli in the current condition.

The aim of the baseline measurement was to find out how long it took participants on average to respond to simple stimuli of varying durations but always devoid of finality cues, presented in a certain modality. The participants' task was to press a designated button as soon as the end of the stimulus was reached. In the audio-visual modality, the baseline stimuli consisted of a video still (a single frame of some speaker) 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, in the audio-only baseline measurement, only the stationary /m/ was heard.

During the actual end-of-utterance detection part, participants were given a dual task: a prediction task and a monitoring task. For the prediction task, they had to indicate, as soon as possible, when the speaker finished his or her utterance by pressing a dedicated button. For the monitoring task, participants had to press another button as soon as they saw a red dot appear on the screen. These red dots were added to a limited number of dummy stimuli to make sure that participants paid attention to the visual information on the screen. 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

FIG. 1. 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.”

duration of the red dot appearance was 1/25s (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.

TABLE I. Reaction times in milliseconds for the different conditions (AV, audio-visual; VO, vision-only; AO, audio-only) in both the baseline measurement and the actual experiment. In the latter case, the results are also split with respect to length.

			Length	
Measurement	Condition	RT	3 words	5 words
Baseline	AV	392		
	VO	331		
	AO	380		
Experiment	AV	509	585	433
	VO	669	804	533
	AO	525	628	422

4. Data processing

Reaction times (RT) 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 1000ms after the speaker finished speaking, 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. In the case of the baseline measurement we can be certain that these are errors, and hence they were replaced by the mean RT value for that stimulus. In the actual end-of-utterance experiment a negative RT is not necessarily an error (because the participant may know the end is near even though the speaker has not actually stopped speaking), and hence these were not removed. It is important to note that these negative RTs were very rare and their inclusion did not significantly alter the results. There was a total of 23 non-responses (0.4%), 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). Where the sphericity assumption was violated, we used the Greenhouse-Geisser correction on the degrees of freedom. Post hoc analyses were performed with the Bonferroni method.

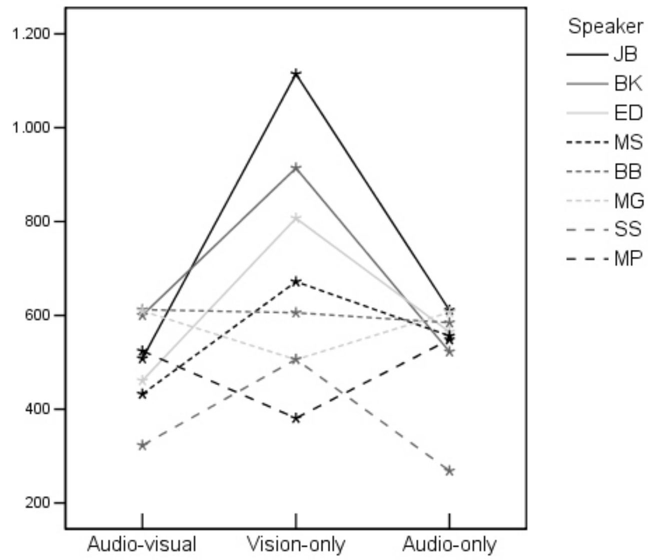

FIG. 2. The mean reaction time (in ms.) for the different speakers in the three modalities.

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 vision-only (VO) condition evoked the fastest reaction times followed by the audio-only (AO) and the audio-visual (AV) conditions. An ANOVA 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 ($F(2,58) = 11.215, p < .001$). Post hoc analyses showed that there was a significant difference between the audio-visual and vision-only condition ($p < .001$), and between the vision-only and the audio-only condition ($p < .001$). The audio-only and the audio-visual condition did not, however, differ significantly ($p = .368$). The stimuli in the baseline variant differed in duration, but this did not have a significant influence on the reaction times ($F(7,203) = 2.891, p = .065$), nor was the interaction between condition and stimulus duration significant ($F(14,406) = 2.021, p = .14$).

Next consider the results of the actual experiment. Here the audio-visual (AV) condition yielded the quickest responses, followed by the audio-only (AO) condition, while the vision-only (VO) condition lead the slowest reaction times. An ANOVA with condition, length (measured by the number of words: 3 or 5), and speaker as within participants variables and reaction time as the dependent variable was carried out. A significant main effect of condition was found ($F(2,58) = 17.052, p < .001$). Post hoc analyses showed that there was a significant difference between the audio-visual and vision-only condition ($p < .001$), and between the vision-only and the audio-only condition ($p < .001$). The audio-only and the audio-visual condition did not differ significantly. In addition, a main effect of stimulus length was found ($F(1,29) = 90.086, p < .001$). Inspection of Table I reveals that 3 word utterances led to longer reaction times than 5 word utterances, perhaps because shorter answers may contain less cues for making a correct prediction. Finally, there was also a main effect of speaker ($F(7,203) = 23.500, p < .001$) 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 ($F(2,58) = 26.480, p < .001$) was found. As can be seen in Table I, the difference between the RT for 3 word utterances and for 5 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. As can be seen in Figure 2, the RT patterns for the different speakers in the different conditions mostly follow the same pattern, but some speakers score particularly good in one of the conditions, for instance, because they better cue the end of their utterance using facial cues rather than auditory ones. The remaining significant interactions all involve the factor speaker and are not separately discussed.

It is highly 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 to the reaction times in the actual experiment, in order to eliminate the influence of the presentation modality itself. The picture that emerges is visualised in Figure 3, which shows that the reaction times for the two sessions are more similar in the audio-visual condition, and more divergent in the vision-only condition, while the results for the audio-only condition are in between 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 difference for significance, we performed a univariate ANOVA with average RT for each participant as dependent variable, and experimental variant (baseline measurement versus actual experiment) and condition (AV, AO, VO) as independent variables, which indeed showed a significant 2-way interaction between these two factors ($F(2,174) = 12.106, p < .001$).

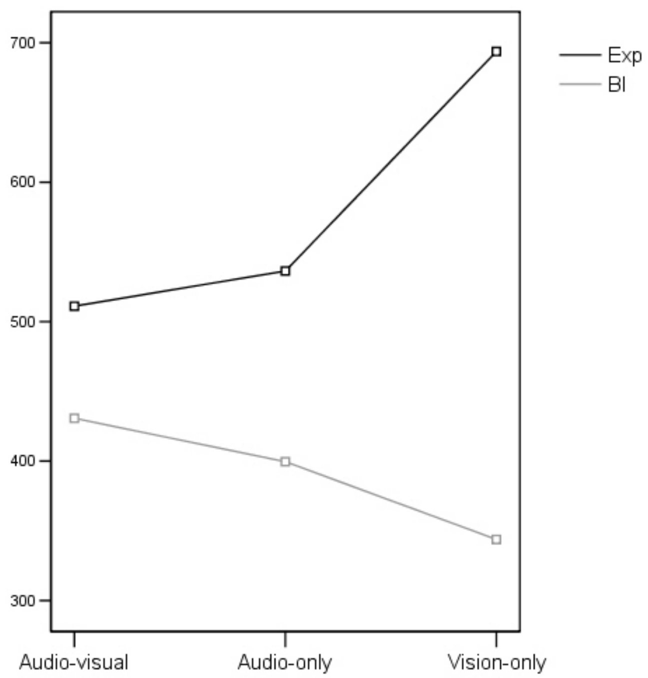

FIG. 3. The mean reaction time (ms) in the three conditions for the baseline and the actual experiment.

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 were impossible, the reaction times for the different modalities in the actual experiment would have been the same, or at least have the same patterns 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, followed by the audio-only stimuli and the vision-only stimuli. This already suggests that combining modalities is useful for end-of-utterance detection, but the differences with the audio-only stimuli were relatively small and, in addition, it is unclear to what extent cues in the different modalities help in detecting end of an utterance. For instance, the contribution of vision-only cues for end-of-utterance detection is somewhat unclear in the first experiment. These issues are investigated in a second experiment, where participants have to decide on the basis of small fragments whether these fragments marked the end of an utterance or not. As before, participants have to perform this task in two unimodal (audio-only, vision-only) and one bimodal condition (audio-visual).

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 hear 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 were selected from the utterances of the same 8 speakers which were used in experiment I. For each of these speakers we randomly extracted answers from their original set of answers (see section II), and constructed two types of fragments from these: short ones, consisting of one word, and long ones, consisting of two words. Half of the fragments were from a final (end-of-utterance) and half were taken from a non-final position.

For each of the eight speakers, we created 4 short pairs (final/non-final) and 4 long pairs of fragments, where the short fragments always consisted of the last word of the corresponding long (two word) fragment. The length of the original context surrounding a fragment was more or less balanced, with a small majority of fragments extracted from answers containing longer (five word) lists. To guarantee the understandability of the fragments and to make sure they are 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 non-final 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 exactly as long. As for experiment I, all fragments were stored in three ways: audio-only (AO), vision-only (VO) or audio-visually (AV).

Therefore, in total 128 stimuli were created for each modality: 8 speakers $\times$ 2 lengths (short-long) $\times$ 2 types (non-final and final) $\times$ 4 instances.

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.

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. Experiment II had a counterbalanced within-participants design, consisting of 3 conditions, one containing audio-visual (AV), one audio-only (AO) and one vision-only (VO) stimuli, presented to participants in the same way as in experiment I. The order in which participants saw the three conditions was systematically varied.

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 in order to minimize possible learning effects. Participants were exposed to either the A-versions or the B-versions of a list. So, each participant passed the items in a different random order in each part. Each condition was preceded by a short practice session, consisting of two stimuli, so that participants could get used to the type of tasks and stimuli.

4. Statistical analyses

All tests for significance were performed with a multinomial logistic regression, with as dependent variable the percentage of correct classifications (i.e., final and non-final stimuli classified as final and non-final respectively).

B. Results

Table II gives the overall results for three factors of interest, i.e., fragment type, stimulus length and modality. According to the multinomial logistic regression all three factors had a significant influence on the classification. First, consider the main effect of fragment type. It appears that judging non-finality is somewhat easier than judging finality (80.8 vs. 75.2 percent), but overall it is clear that the vast majority of the fragments is classified correctly. Stimulus length also had a significant influence, as can be seen in Table II, with short (one word) fragments being somewhat more difficult than longer (two word) fragments. The most interesting main effect is that of modality. It is interesting to note that both unimodal conditions yield around 75% correct (75.7 for the vision-only condition and 73.7 for the audio-only condition), and that both are clearly outperformed by the bimodal, audio-visual condition (with almost 85% correct). This pattern of results is visualized in Figure 4.

TABLE II. For each factor, the levels of the factor, the percentage of correct judged utterances, and the multinomial logistic regression statistics are given.

Factor	Level	% Correct	MLR statistics
Fragment type	NF	80,8	$\chi^2(1) = 35.073, p < .001$
	F	75.2	
Stimulus length	Short	75.1	$\chi^2(1) = 39.185, p < .001$
	Long	81.0	
Modality	AV	84.7	$\chi^2(2) = 108.245, p < .001$
	VO	75.7	
	AO	73.7	

TABLE III. For each speaker, the total percentage of correct judged utterances, and the percentage of correct judged utterances as a function of the 3 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

Besides the main effects for the three factors listed in Table II, the factor speaker also had a significant main effect ($\chi^2 = 276.887, df = 7, p < .001$). As can be seen in Table III, the total number of correct classifications differs per speaker, ranging from 63% correct for speaker JB to 87.8% for speaker SS. This shows that there are overall substantial differences between speakers in end-of-utterance signalling.

It is rather interesting to observe that the scores per speaker may differ across conditions. Indeed, a significant 2-way interaction was found between speaker and modality ($\chi^2 = 174,061, df = 14, p < .001$); in Table III it can be seen that, for instance, speaker BB apparently offers clearer visual than auditory cues, as the percentage of correctly classified stimuli drops considerably in the AO condition. This is different for speaker MG, for instance, who seems to send more useful auditory (in her case the classification scores drop in the VO condition).

Another significant 2-way interaction was found between fragment type and modality ($\chi^2 = 181.402, df = 4, p < .001$). Table IV illustrates this interaction. It can be seen that both for the non-final and final fragments, the number of correctly classified audio-visual stimuli is about equally high (85.6% and 83.8%), but the unimodal conditions (VO and AO) score relatively better for the non-final than for the final fragments.

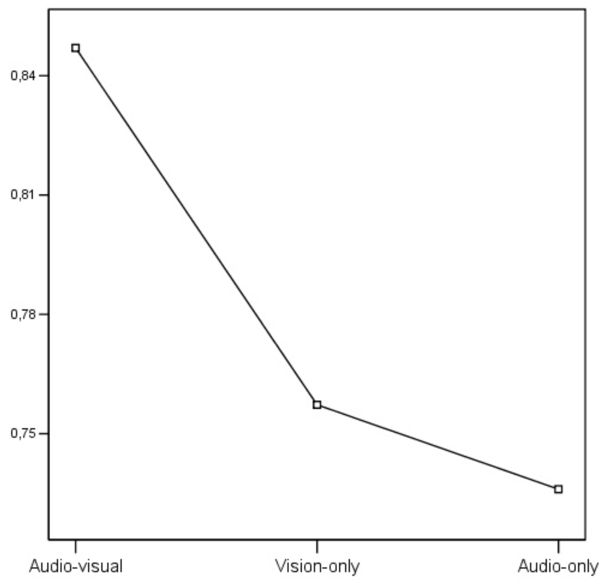

FIG. 4. Percentage of correct answers in the audio-visual (AV), vision-only (VO) and audio-only (AO) conditions.

TABLE IV. For each stimulus modality and fragment type, the percentage of correct judged utterances per fragment type

Response	AV	VO	AO	Total
NF	85.6	79.4	77.4	80.8
F	83.8	72.1	69.8	75.2
Total	84.5	75.3	72.7	

Moreover, a significant two-way interaction was found between fragment type and stimulus length ($\chi^2 = 181.402, df = 4, p < .001$). This interaction can be explained by looking at Table V, where it can be seen that for the non-final 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% resp.).

Table V also illustrates a second, significant 2-way interaction, between stimulus length and modality ($\chi^2 = 181.402, df = 4, p < .001$). 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%). 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%).

Finally, a significant 3-way interaction was found between stimulus length, fragment type and modality ($\chi^2 = 223.792, df = 11, p < .001$). Inspection of Table V reveals that this interaction can be explained as follows: for the short utterances, the differences between non-final and final correctness scores in the 3 different modalities are always roughly the same, however, when looking at the long utterances, it can be seen that there is a sizeable gap between the scores for final and non-final stimuli for the unimodal conditions.

TABLE V. For each modality, the percentage of correct judged utterances, as a function of stimulus length (1 or 2 words) and fragment type (non-final and final).

Length	Response	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		84.5	75.3	72.7	

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 lowest scores are obtained with the AO condition, which has received 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. The non-final fragments were slightly more often judged correctly than the final fragments. For the non-final 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. Various researchers have pointed out that speakers may also employ visual cues (such as posture, head movements or gaze) for this purpose. This naturally raises the question which modalities people employ to determine when a speaker is at the end of an utterance, which is the central question addressed in this paper. In order to answer this question, we first collected utterances in a semi-spontaneous way using a new experimental paradigm eliciting target list-answers. On the basis of these utterances, two perception experiments were carried out.

The first experiment was a reaction time experiment in which participants were confronted with utterances, taken out of their original interview context, and presented in three formats: vision-only, audio-only or audio-visual. 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 margin) by the audio-only condition, and with the slowest responses in the vision-only condition. It is interesting to observe that the results were the exact mirror image of a baseline reaction measurement, where participants had to react to stimuli without finality cues.

This difference in response to cue-bearing and cue-less stimuli can 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, 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)).

A second experiment was conducted because the differences in reaction times between the audio-visual and audio-only conditions were relatively small, and to get more insight in how participants respond to stimuli in the different modalities. In this experiment participants were offered short (1 word) and long (2 word) fragments which either did or did not mark the end of an utterance, and participants had to classify these as final or non-final. Again, the bimodal presentation format gave the best results: 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 are made in the audio-only condition, i.e., the situation which has received most attention in the literature so far. Two possible explanations for the superiority of the audio-visual stimuli exist. 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 other more auditory ones. Clearly, this also speaks in favour of a bimodal presentation.

In addition, it was found that non-final fragments were somewhat more often classified correctly than the final ones. And for the non-final 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 for the final fragments. This suggests that when finality cues are available, it makes no difference whether the fragment is short or long, but when no finality cues are 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, while when no finality is displayed, people are searching for more global cues. It could also be caused by the fact the finality is marked by the presence and absence of one or more marked features, and that people spot finality by looking for the presence of that cue. It may be just more easy to see whether a cue is present than to decide that something is not there (Hearst (1981)).

Related to this, it is interesting to observe that the difference between short and long fragments in the 2nd experiment has more impact on the audio-only stimuli than on the vision-only stimuli. That in the audio-only condition the longer fragments are better classified than the short fragments suggests that the finality cues in speech are more global ones, and hence that participants can make better judgements for longer fragement 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 of the fact that the audio-only condition outperforms the vision-only condition in experiment 1, but not in experiment 2. Since the stimuli in the second experiment where overall shorter fragments (consisting of 1 or 2 words) than those in the first experiment (which consisted of entire utterances of 3 or more words), the participants in the second experiment could not use the spoken global cues to full effect.

VI. ACKNOWLEDGEMENTS

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.

VII. REFERENCES

Argyle, M., & Cook, M. (**1976**), *Gaze and mutual gaze.* Cambridge: Cambridge University Press.

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

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

Caspers, J. (**1998**). Who’s next? The Melodic Marking of Questions vs. Continuation in Dutch. *Language and Speech*, **41**, 375-398.

Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., & Rich, C. (**2001**). Non-Verbal Cues for Discourse Structure.

Proc. of the 41st Annual Meeting of the Association of Computational Linguistics (ACL), Toulouse, France.
Couper-Kuhlen, E. (**1993**). *English speech rhythm.* Philadelphia: Benjamins.
Doherty-Sneddon, G., Bonner, L., & Bruce, V. (**2001**). Cognitive demands of face monitoring: Evidence for visuospatial overload. *Memory & Cognition*, **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. *Optometry and Vision Science*, **78**, 712-725.
Duncan, S. (**1972**). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, **23**, 283-292.
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: Cambridge University Press, pp. 169-202.
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, *Language and Cognitive Processes*, **11**, 597-604.
Hearst, E. (**1991**). Psychology and nothing, *American Psychologist*, **79**, 432-443.
Kendon, A. (**1967**). Some functions of gaze-direction in social interaction. *Acta Psychologica*, **26**, 22-63.
Kim, J., Davis, C., & Krins, P. (**2004**). Amodal processing of visual speech as revealed by priming. *Cognition*, **93**, B39-B47.
Kobayashi, H., & Kohshima, S. (**1997**). Unique morphology of the human eye. *Nature*, **387**, 767 - 768.
Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (**1998**). An Analysis of Turn-Taking and Backchannels Based on Prosodic and Syntactic Features in Japanese Map Task Dialogs. *Language and Speech*, **41**, 295 - 321.
Krahmer, E. & Swerts, M. (**2004**). More about brows. In: *From brows to trust: evaluating embodied conversational agents*, C. Pelachaud & Zs. Ruttkay (eds.), Dordrecht: Kluwer, pp. 191-216.
Leroy, L. (**1984**). The psychological reality of fundamental frequency declination. *Antwerp papers in linguistics*, **40**, Antwerp University.
Levinson, S. (**1983**). *Pragmatics.* Cambridge: Cambridge University Press.
Maynard, S. K. (**1987**). Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation. *Journal of Pragmatics*, **11**, 589-606.
Nakano, Y. I., Reinstein, G., Stocky, T., & Cassell, J. (**2003**). Towards a Model of Face-to-Face Grounding. *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan.
Novick, D. G., Hansen, B., & Ward, K. (**1996**). Coordinating turn-taking with gaze. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, PA.
de Pijper, J. R. & Sanderman, A. A. (**1994**). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *The Journal of the Acoustical Society of America*, **96 (4)**, 2037-2047.
Price, P., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, S. (**1991**). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, **90**, 2956-70.
Schwartz, J.-L., Berthommier, F., & Savariaux, C. (**2004**). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, **93**, B69-B78.
Silverman, S. & Pierrehumbert, J. (**1990**). The timing of prenuclear high accents in English. In: *Laboratory Phonology, Vol I: Between the grammar and physics of speech*, J. Kingston and M. Beckman (eds.), pp. 71-106, Cambridge: Cambridge University Press.
Swerts, M. (**1997**). Prosodic features at discourse boundaries of different strength. *The journal of the Acoustical Society of America*, **101**, 514-521.
Swerts, M. (**1998**). Filled pauses as markers of discourse structure. *Journal of pragmatics*, **30**, 485-496.
Swerts, M., Bouwhuis, D., & Collier, R. (**1994**). Melodic cues to the perceived finality of utterances. *Journal of the Acoustical Society of America*, **96**, 2064-2075
Swerts, M., Collier, R., & Terken, J. (**1994**). Prosodic predictors of discourse finality in spontaneous monologues. *Speech Communication*, **15**, 79-90.
Swerts, M., & Geluykens, R. (**1994**). Prosody as a marker of information flow in spoken discourse. *Language and Speech*, **37**, 21-43.
Vertegaal, R., Slagter, R., van de Veer, G., & Nijholt, A. (**2000**). Why conversational agents should catch the eye. *Proceedings of the Internation Computer-Human Interaction conference (CHI)*, The Hague, The Netherlands.
Walley, A. C., Michela, V. & Wood, D. (**1995**). The gating paradigm: Effects of presentation format on spoken word recognition by children and adults. *Perception & Psychophysics*, **57**, 343-351.
Ward, N. & Tsukahara, W. (**2000**). Prosodic Features which Cue Back-channel Responses in English and Japanese. *Journal of Pragmatics*, **23**, 1177–1207.
Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M. & Price, P. (**1992**). Segmental Durations in the Vicinity of Prosodic Phrase Boundaries. *Journal of the Acoustical Society of America*, **91**, 1707-1717.