Predicting end of utterance in multimodal and unimodal conditions

Pashiera Barkhuysen, Emiel Krahmer, Marc Swerts

Communication and Cognition
Faculty of Arts, Tilburg University
P.O.Box 90153, NL-5000 LE Tilburg, The Netherlands
{p.n.barkhuysen, e.j.krahmer, m.g.j.swerts}@uvt.nl

Abstract

In this paper, we describe a series of perception studies on uni- and multimodal cues to end of utterance. Stimuli were fragments taken from a recorded interview session, consisting of the parts in which speakers provided answers. The answers varied in length and were presented without the preceding question of the interviewer. The subjects had to predict when the speaker would finish his turn, based on video material and/or auditory material. The experiment consisted of 3 conditions: in one condition, the stimuli were presented as they were recorded (both audio and vision), in the two remaining conditions stimuli were presented in only the auditory or the visual channel. Results show that the audiovisual condition evoked the fastest reaction times and the visual condition the slowest. Arguably, the combination of cues from different modalities function as complementary sources and might thus improve prediction.

1. Introduction

In order to smoothly switch turns during a conversation, dialogue participants make use of a turn taking mechanism so as to avoid simultaneous talking (Beattie, 1982; Duncan, 1972). Due to cognitive limitations in working memory people find it hard to speak and listen at the same time (Beattie, 1982). The turn-taking system can be described as a set of rules, such as the rule that in principle only one speaker speaks at the same time (Sacks, Schegloff, & Jefferson, 1972). When the current speaker finishes a turn, another participant may take over, either because the current speaker selects the next speaker or by self-selection (Duncan, 1972; Sacks, Schegloff, & Jefferson, 1972). In a similar vein, a speaker who wishes to avoid being interrupted while still speaking, must display cues signalling that he is currently not yet willing to give up the floor (Beattie, 1982).

There is ample empirical evidence that speakers use cues to signal when they are reaching the end of their current utterance and that listeners are able to detect such end-of-utterance cues. These cues may be auditory ones, such as intonation (e.g., Swerts, Bouwhuis, & Collier, 1994; Caspers, 1998; Koiso, Horiuchi, Tutiya, Ichikawa, & Den, 1998). End-of-utterance marking can also be signaled by visual cues, such as gestures and postural shifts (Beattie, 1982; Cassell, Nakano, Bickmore, Sidner, & Rich, 2001; Duncan, 1972). with special attention to the function of gaze (Argyle & Cook, 1976; Kendon, 1967). Other facial cues are eyebrow movements (Cavé, Guítella, Bertrand, Santi, Harlay, & Espesser, 1996), head movements (Maynard, 1987), and blinking (Doughty, 2001).

This raises the natural question: how do these visual cues relate to the auditory ones? Are listeners more sensitive to auditory or to visual cues, or do they use these cues most efficiently when both groups of cues are present? In this paper, we report on a reaction time experiment with the intention to determine the relative weight of the different modalities used for end-of-utterance marking. We compare a multimodal condition in which subjects have both auditory and visual cues at their disposal (stimuli presented as they were produced) with two unimodal conditions where subjects could only use auditory or visual cues.

2. General procedure

2.1. Data collection

As stimuli for our reaction time experiment, we used recordings of speakers who had to respond to questions in an interview situation. These questions were intended to evoke lists of words, such as questions requiring general knowledge, like "What are the colors of the Dutch flag" or questions eliciting a set of numbers, such as "What are the odd numbers between five and fifteen?". The lists that were asked for varied in length, consisting of sequences of 3 to 5 words. A total of 22 speakers was filmed (13 male and 9 female). The interview consisted of 33 questions, of which 25 were experimental and 8 were filler items, i.e. questions where the number of words in the answers could in principle not be predicted, like "What languages do you speak?". The original recordings were made with a digital video camera (25 frames per second). The recordings were read into a computer and orthographically transcribed.

2.2. Selection of stimuli

The stimuli were randomly selected from the transcriptions of 8 speakers (4 male and 4 female), and consisted of speakers' answers without the preceding question of the interviewer. Each stimulus continued for 1000ms after the speaker finished speaking. Per speaker, 3 instances of answers consisting of 3 words and 3 instances of 5 words were selected. In addition, for each speaker 2 filler items were selected consisting of 1, 4 or more than 5 words, or including other spoken text (such as repetitions of the question or fragments where speakers think aloud).

2.3. Subjects

A group of 30 subjects (7 male and 23 female) participated; all were native speakers of Dutch, and 8 were left-handed. The subjects were between 24 and 62 years old. None of the subjects had participated as a speaker in the data collection phase.

2.4. Procedure

The experiment consisted of 3 conditions, in which the stimuli were presented audio-visually (AV), audio-only (AO) and vision-only (VO). In the AV condition, subjects saw the stimuli

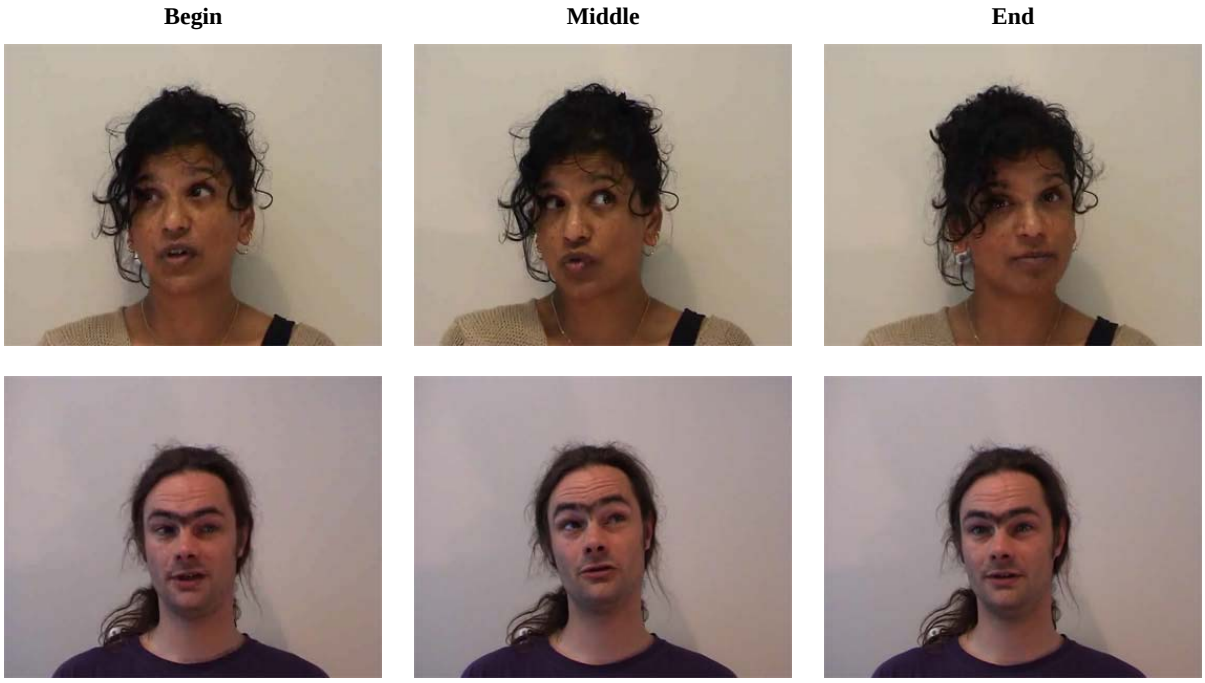

Figure 1: The speakers SS and BB while uttering the first and middle word and just after the final word of a three word answer, such as "red, blue, white."

as they were recorded, in the AO condition subjects heard the speaker while the visual channel only depicted a static black screen, and in the VO condition subjects only saw the speaker. All subjects participated in all three conditions; the order in which subjects passed these conditions differed (counterbalanced, within subjects design). Within a condition, the 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 subjects acquainted with the experimental setting and the stimuli.

During the baseline measurement subjects were confronted with stimuli of different durations and devoid of finality cues. Their task was to press a designated button as soon as the end of the stimulus was reached. In the AV condition, the baseline stimuli consisted of a video still (a single frame of one of the speakers) accompanied with 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 VO baseline condition, only the video still was displayed, in the AO baseline condition, only the stationary /m/ was heard. The aim of the baseline session was to find out how long it took subjects on average to respond to a simple stimulus presented in a certain modality and to control for inter-individual differences.

During the actual end-of-utterance detection part, subjects 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, again by pressing a dedicated button at this exact moment. For the monitoring task, subjects 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 subjects in the audiovisual condition listened to *and* watched the stimuli. The duration of the red dot appearance was 1/25s (a single frame); it appeared at varying locations on the screen. These dummy stimuli were not used in further analyses.

Table 1: Reaction time in milliseconds for the different conditions in the baseline session and the experiment

	Baseline	Experiment			Diff
	Total	3wrd	5wrd	Total	BL-Exp
AV	430,713	585,000	432,514	508,757	117,011
VO	343,817	828,611	547,821	688,216	344,399
AO	399,567	637,581	427,921	532,751	133,184

2.5. Design

The baseline session had a 3 (condition) x 8 (speakers) factorial design, with condition and speaker as within subjects variables. The experiment had a 3 (condition) x 8 (speakers) x 2 (words) factorial design, with condition and speaker and number of words (3 or 5) as within subjects variables.

3. Results

3.1. Baseline

A first inspection of the baseline measurements revealed that a number of reaction times were far below or above the mean value for a certain stimulus. As the task was to press the button immediately after the stimulus stopped, we removed all values below 0 ms and replaced them by the mean value for that stimulus. Next, we removed all extreme values as indicated by the box plot for that stimulus and replaced them also by the mean. In the baseline session, the visual presentation mode evoked the fastest reaction times ($M = 343,817$) followed by the auditory condition ($M = 399,567$), and the audiovisual condition

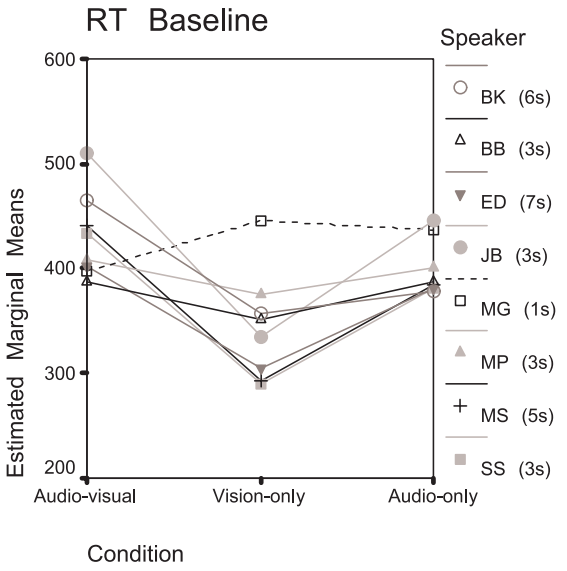

Figure 2: The mean reaction time for the different groups of stimuli in the baseline session split by the three presentation modes. (Durations in seconds between brackets.)

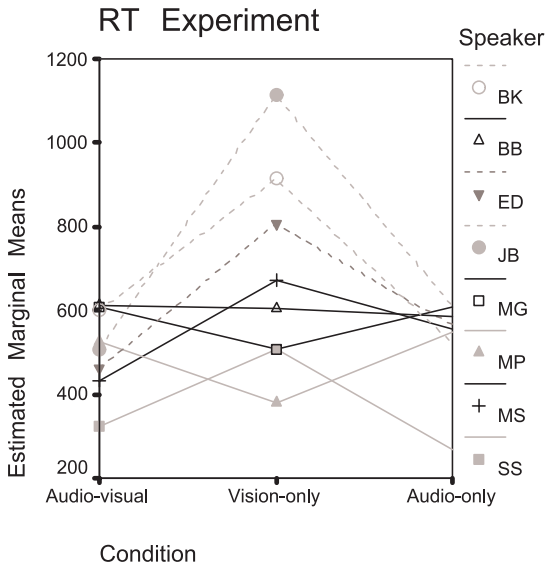

Figure 3: The mean reaction time for the different groups of stimuli in the actual experiment split by the three presentation modes.

$(M = 430, 713)$. These results can be found in Table 1. Differences were tested for significance by a repeated measurements analysis of variance, with speakers and condition as within subjects variables. There was a significant difference between the three conditions $(F(2, 58) = 11, 215; p < .001)$. Post hoc analyses with the Bonferroni-method showed that there was a significant difference between the audiovisual and vision-only condition $(p < .002)$, and between the vision-only condition and the auditory condition $(p < .001)$. The auditory condition and the audiovisual condition did not, however, differ significantly $(p = .368)$. The mean reaction times as a response to a different speaker are plotted for the three conditions in Figure 1. When looking at speakers individually, the general pattern is that the visual condition evokes the fastest RT's. Most speakers evoke this same pattern in their responses, although speaker MG shows a reversed pattern. The difference between MG and the other speakers is statistically significant $(F(14, 406) = 2, 021; p = .015)$.

3.2. Experiment

In the actual experiment, the audiovisual presentation mode evoked the fastest reaction times $(M = 508, 757)$ followed by the auditory condition $(M = 532, 751)$. Here, the visual condition was the slowest $(M = 688, 216)$. Differences were tested for significance by a repeated measurements analysis of variance, with speakers, condition and number of words as within subjects variables. There was a significant difference between the three conditions $(F(2, 58) = 17, 052; p < .001)$. Post hoc analyses with the Bonferroni-method showed that there was a significant difference between the audiovisual and vision-only condition $(p < .001)$, and between the vision-only condition and the auditory condition $(p < .001)$. The auditory condition and the audiovisual condition did not, however, differ significantly $(p = 1.0)$. In Table 1, the results are split for the 3-word condition and the 5-word condition. The reaction times for the 5-word condition are faster than for the 3-word condition. A possible explanation is that when the given answer is longer, more cues are available to make a good prediction. The effect that the visual condition is slower is stronger in the 3-word condition than in the 5-word condition, which is significant $(F(2, 58) = 4, 133; p = .021)$. In Figure 2, the reaction times per speaker are displayed per condition. Most speakers show the same pattern as discussed in Table 1, i.e. that the visual condition is the slowest, but the speakers BB, MG and MP show a different, reversed pattern.

3.3. Combined picture

The picture that emerges from the two sets of results presented above is that the reaction times in the baseline condition are essentially different from those obtained in the actual experiment. That is, where the vision-only condition leads to the fastest RT results in the baseline condition, they are the slowest in the actual experiment. The reverse is true for the data in the audiovisual condition, whereas the data for the audio-only condition are in the middle in both sessions. A univariate ANOVA with average RT for each subject as dependent variable, and experiment (baseline versus actual experiment) and modality (AV, AO, VO) as independent variables indeed showed a significant 2-way interaction between these two factors on the reaction times $(F(2, 174) = 12, 106; p < .001)$. The pattern of results is visualised in Figure 4, which shows that the reaction times for the two sessions are more similar in the audiovisual condition, and very divergent in the vision-only condition, where the results for the audio-only condition are in between these two extremes.

4. Discussion

The present study tested whether listeners can predict the end of an utterance, and what the cue value is of visual prosodic cues versus auditory prosodic cues. It was found that the audiovisual presentation mode evoked the fastest reaction times,

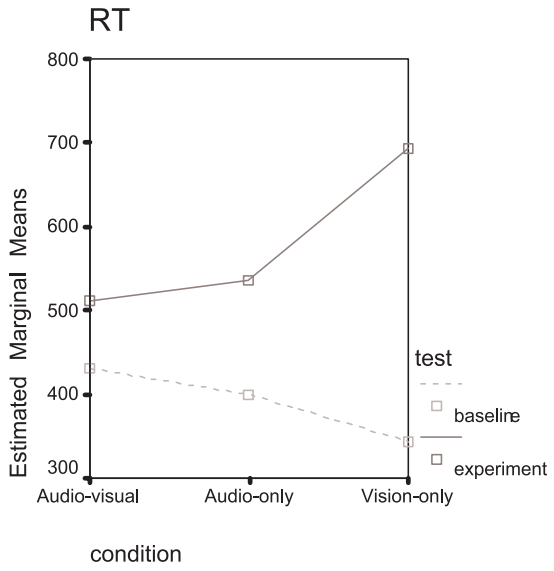

Figure 4: Mean RT results for 2 experiments (baseline session=bottom line, experiment=top line) in three conditions.

as opposed to the visual mode. This implies that prediction of the end of an utterance improves with the increase of cues from different modalities. If prediction of the end of a turn were impossible, the reaction times in the different modalities would have been the same, or at least have the same patterns as in the baseline session, where no cues were present. In the baseline session, however, the opposite effect was found. Here the visual condition evoked the fastest reaction times, as opposed to the audiovisual condition, in which the two modalities are combined. These apparent contradictory results 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, Bonner, & Bruce, 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 (Schwartz, Berthommier, & Savariaux, 2004). There is evidence that auditory speech can indeed be compared with visual speech in this way. Kim, Davis & Krins (2004) suggest that the same or alike processes are involved in the processing of visual versus auditory speech, as they found that visual speech primes can be used with targets presented in different modalities. Thus, we suppose that the difference in reaction times is a reflection of the cue value of the different modalities.

5. 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). Marc Swerts is also affiliated with Antwerp University and with the FWO-Flanders. Thanks are due to Carel van Wijk for statistical advice, Lennard van de Laar for technical assistance and Jean Vroomen for allowing us to make use of the Pamar software.

6. References

[1] Argyle, M., & Cook, M. (1976). Gaze as part of the sequence of interaction, *Gaze and mutual gaze* (pp. 98-124). Cambridge: Cambridge University Press.

[2] Beattie, G. W. (1982). Turn-taking and interruption in political interviews: Margareth Thatcher and Jim Callaghan compared and contrasted. *Semiotica*, **39** (1/2), 93-114.

[3] Caspers, J. (1998). Who's next? The melodic marking of question vs. continuation in Dutch. *Language and Speech* **41** (3-4), 375-398.

[4] Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., & Rich, C. (2001). *Non-Verbal Cues for Discourse Structure*. Paper presented at the Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics, Toulouse, France.

[5] Cavé, C., Guaítella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). *About the relationship between eyebrow movements and f0 variations*. Paper presented at the Proceedings of the ICSLP, Philadelphia.

[6] Doherty-Sneddon, G., Bonner, L., & Bruce, V. (2001). Cognitive demands of face monitoring: Evidence for visuospatial overload. *Memory & Cognition*, **29** (7), 909-917.

[7] 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** (10), 712-725.

[8] Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, **23** (2), 283-292.

[9] Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, **26**, 22-63.

[10] Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition*, **93** (1), B39-B47.

[11] 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** (3-4), 323-350.

[12] Maynard, S. K. (1987). Interactional functions of a nonverbal sign. Head movement in japanese dyadic casual conversation. *Journal of Pragmatics*, **11** (5), 589-606.

[13] Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language and Speech*, **50**, 696-735.

[14] Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, **93** (2), B69-B78.

[15] Swerts, M., Bouwhuis, D.G., & Collier, R. (1994). Melodic cues to the perceived finality of utterances. *Journal of the Acoustical Society of America*, **94** (4), 2064-2075.