How Auditory and Visual Prosody is Used in End-of-Utterance Detection

Pashiera Barkhuysen, Emiel Krahmer, Marc Swerts

Communication & Cognition
Tilburg University, 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 using visual and auditory cues to end-of-utterance. Fragments were taken from a recorded interview session, consisting of the parts in which speakers provided answers. Final and non-final parts of these fragments were used, varying in length. The subjects had to assess whether the speaker had finished his or her turn, based upon these fragments. The fragments were presented in 3 modalities: either a bimodal presentation mode (both auditory and visually), or in only the auditory or the visual mode. Results show that the audio-visual condition evoked the highest proportion of correct classifications and the auditory condition the lowest. Thus, the combination of modalities clearly works best. Also, non-final fragments are classified better than final ones, and longer fragments are classified better than short ones. It furthermore appears that these factors are different for different modalities: longer fragments are better classified in the auditory modality, while for short fragments the visual modality works better. This suggests that people may make more use of global cues in the auditory modality, while for the visual modality local cues are sufficient.

Index Terms: Audiovisual speech, prosody, end-of-utterance detection, speech production, speech perception.

1. Introduction

Speakers send signals to listeners about the status of their turn, and indicate when they are going to finish speaking, in order to smoothly hand over their turn. In order to do so, they make use of a turn-taking mechanism, which has been described elaborately in the literature [7]. This mechanism works so well that overlaps during turn-transitions rarely occur [7]. Therefore, listeners must somehow be able to detect such end-of-utterance cues. Previous studies report which possible auditory cues such as intonation, rhythm and pause, or which visual cues such as facial expressions, a speaker's gaze, gestures, or postural shifts may function as turn-transition signals [8, 9, 10, 12, 17].

A lot of research has been devoted to the *auditory modality* [3, 8, 17]. [3], to give one example, found that subjects who couldn't understand Swedish, predicted upcoming prosodic boundaries in spontaneous Swedish speech just as well as native Swedish speakers. Their judgments were influenced by factors such as the presence/absence of final creak and phrase-final f0 level and slope [3]. Cues operating in the *visual modality* during a dialogue are gestures and postural shifts [4] and gaze [1, 9, 13, 15]. [1] describes specific gaze patterns during turn-taking. Speakers tend to stare at the listener when they start a new turn, while listeners look away immediately after they have handed over the previous turn. Later in the turn, listeners look at speakers, and when the speaker looks at the listener during a pause, there is a considerable chance that the listener will take over the turn again. Head movements and eyebrow movements seem to have a variety of functions, including end-of-utterance detection [5, 14, 16]. Blinking can also play a role in a dialogue, because the blinking rate increases when people are involved in a conversation, compared with other situations, although it is yet unclear what that precise role can be [6].

Although there is much report about which individual behavioral expressions may function as important cues in end-of-utterance marking, less is known about the relative weight of the auditory and the visual modality. Therefore, the aim of this paper is to get more insight in the relative contribution of these two modalities in the turn-taking mechanism, and to explore how sensitive listeners are to signals displayed in different modalities. In an earlier study, we conducted a reaction time experiment [2] where we compared two conditions in which participants had only cues in one modality at their disposal (film fragments presented without their original corresponding sound or vice-versa) with a third, bimodal condition in which participants could use both auditory and visual cues. The audio-visual stimuli led to the quickest responses, followed by the audio-only stimuli and the vision-only stimuli. This suggested that combining modalities is useful for end-of-utterance detection, but the differences between the bimodal and the unimodal, audio-only condition were relatively small. In addition, it remained unclear to what extent the contribution of vision-only cues can help in end-of-utterance detection. Are visual expressions a supplement or can facial expressions have an even stronger effect than intonation alone? Is it possible to detect that someone has finished speaking solely on the basis of facial expressions? We are also interested to see how more local cues in different modalities contribute to end-of-utterance detection.

These issues are further investigated in the current paper. We set up a decision task in which participants have to decide in a binary way on the basis of small fragments whether these fragments mark the end of an utterance or not. We used fragments taken from the same material as in [2] and presented them in the same 3 conditions. 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. 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 [11]. Participants must try to recognize the entire spoken stimulus on the basis of the fragment. 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

Figure 1: *The speakers MP and MS while uttering a non-final and a final word.*

fragment originated (i.e. final or not final).

2. 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 questions eliciting 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 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 not be predicted (e.g., Q: What 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 is 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.

3. Method

3.1. Stimuli

For the current experiment 4 male and 4 female speakers were randomly selected from the corpus of 22 speakers described above. For each of these speakers we randomly extracted answers from their original set of answers (see section 2 - Audio-visual recordings), and constructed two types of fragments from these: short ones, consisting of 1 word, and long ones, consisting of 2 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 (2-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 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 1-word and 2-word stimuli lasted exactly as long. Like in [2], 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 and long) $\times$ 2 types (final and non-final) $\times$ 4 instances.

3.2. 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 the experiment of [2].

3.3. 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. The experiment had a counterbalanced within-subjects design, consisting of 3 conditions, one containing audio-visual (AV), one audio-only (AO) and one vision-only (VO) stimuli. The order in which participants saw the three conditions was systematically varied.

Each condition consisted of two parts: one part for the short (1-word) fragments and one part for the long (2-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, and to prevent that a non-final and a final fragment of the same speaker are being presented successively. 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.

3.4. Statistical analyses

All tests for significance were performed with a multinomial logistic regression.

4. Results

Table 1 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 1, with short (1-word) fragments being somewhat more difficult than longer (2-word) fragments. The most interesting main effect is that of *modality*. It is interesting to note that both unimodal conditions yield around 75% correct responses (75.7 for the

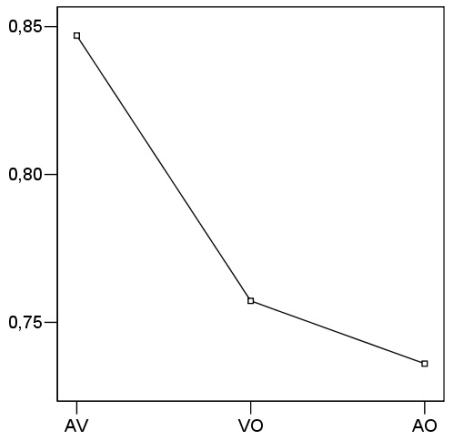

Figure 2: *The proportion of correctly judged utterances is highest in the audio-visual (AV) condition, and is lower in the two unimodal (AO and VO) conditions.*

Table 1: *For each factor, the levels of the factor, the proportion of correctly judged utterances, and the multinomial logistic regression statistics are given.*

Factor	Level	% Correct	χ^2
Fragment Type	Non-Final	80,8	35.073, df = 1,
	Final	75.2	$p < .001$
Stimulus Length	Short	75.1	39.185, df = 1,
	Long	81.0	$p < .001$
Modality	AV	84.7	108.245, df = 2
	VO	75.7	$p < .001$
	AO	73.6	

vision-only condition and 73.6 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 2. Besides the main effects for the three factors listed in Table 1, the factor *speaker* also had a significant main effect ($\chi^2 = 276.887, df = 7, p < .001$). As can be seen in Table 2, 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 2 it can be seen that, for instance, speaker BJ 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 cues (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 3 illustrates this interaction. It can be seen that both for the non-final and final fragments, the number of correctly classified

Table 2: *For each modality, the proportion of correctly judged utterances, as a function of speaker.*

Speaker	AV	VO	AO	All
BJ	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 3: *For each modality, the proportion of correctly judged utterances, as a function of stimulus length (1 or 2 words) and fragment type (Non-Final and Final).*

S.Length	F.Type	AV	VO	AO	All
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	75.1
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	81.0
-	NF	85.6	79.4	77.4	80.8
-	F	83.8	72.1	69.8	75.2
Total		84.7	75.7	73.6	78.0

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.

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 3, 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 3 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 AV modality (82,5% correct answers) is followed by the VO modality (74,9%), and subsequently the AO modality (67.9%). However, for the long fragments, the AV (86.9% correct answers) is followed by the AO modality (79.4%), and subsequently the VO modality (76.6%). Also, within the AO modality the difference between short (67.9%) and long (79.4%) fragments is much larger than in the other two modalities, although in all three conditions the longer fragments perform best.

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 3 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.

5. Concluding remarks

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 audio-only condition, which has received most attention in the literature. The vision-only results are somewhat better, which shows that visual cues are indeed useful for participants for successful end-of-utterance detection, but, as said, the combination of modalities clearly works best. Two possible explanations for this finding exist. First, a combined audio-visual presentation format clearly offers more cues than a single modality (e.g. in ambiguous situations modalities could complement each-other in resolving these issues). Second, 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.

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. This suggests that 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 (i.e. in the last part of a fragment), while when no finality is displayed, people are searching for more global cues. It could also be the case that finality is marked by 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.

In general, the *long* fragments are judged better than short fragments. This could mean that people try to search for more global cues, which they obtain when they can access more information (a longer duration). Pursuing upon this, it is worth noticing that a significant interaction was found between stimulus length and modality. For both stimulus lengths, the audio-visual modality is the easiest one. However, for short fragments the vision-only modality works better, while for long fragments the audio-only modality is more easy. Also, within the auditory modality the difference between long and short fragments is much larger than in the other two modalities. This suggest that people tend to make more use of global cues in the audio-only modality, while for visual end-of-utterance detection local cues are sufficient. However, note that it may still be possible that detecting end-of-utterance is dependent upon the use of global cues when participants are exposed to the whole utterance. Although detecting end-of-utterance in a short, final fragment *must* by definition be based on local cues, that doesn't run out the possibility that participants start to make use of more global cues in the same modality when presented with the whole utterance. In that case it may become possible that the overall results for the audio-only mode become will be than for the vision-only mode.

6. 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). Special thanks goes to Lennard van de Laar for technical assistance and to Carel van Wijk for statistical coaching.

7. 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] Barkhuysen, P. N., Krahmer, E. J., & Swerts, M. G. J. (2005). *Predicting End of Utterance in Multimodal and Unimodal Conditions.* Proc. ICSLP, Lisbon, Portugal.

[3] Carlson, R., Hirschberg, J. & Swerts, M. (2005). Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication*, 46(3-4), 326-333.

[4] Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., & Rich, C. (2001). *Non-Verbal Cues for Discourse Structure*. Proc. ACL, Toulouse, France.

[5] Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). *About the relationship between eyebrow movements and f0 variations*. Proc. ICSLP, Philadelphia, USA.

[6] 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.

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

[8] Ferrer, L., Shriberg, E., & Stolcke, A. (2003). *A prosody-based approach to end-of-utterance detection that does not require speech recognition.* Proc. ICASSP, Hong Kong.

[9] Goodwin, C. (1980). Restarts, Pauses, and the Achievement of a State of Mutual Gaze at Turn-Beginning. *Sociological Inquiry*, 50(3-4), 272-302.

[10] Graf, H. P., Cosatto, E., Strom, V., & Hunag, F. J. (2002). *Visual Prosody: Facial Movements Accompanying Speech*. Proc. FGR, Washington DC, USA.

[11] Grosjean, F. (1996). Gating. *Language and Cognitive Processes*, 11(6), 597-604.

[12] Hadar, U., Steiner, T. J., Grant, E. C., & Rose, F. C. (1984). The timing of shifts of head postures during conversation. *Human Movement Science*, 3, 237-245.

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

[14] McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32, 855-878.

[15] Novick, D. G., Hansen, B., & Ward, K. (1996). *Coordinating turn-taking with gaze*. Proc. ICSLP, Philadelphia, USA.

[16] Stiefelhagen, R., & Zhu, J. (2002). *Head Orientation and Gaze Direction in Meetings.* Proc. CHI, Minneapolis, USA.

[17] Swerts, M., Collier, R., & Terken, J. (1994). Prosodic predictors of discourse finality in spontaneous monologues. *Speech Communication*, 15(1-2), 79-90.