Anticipation of Turn-switching in Auditory-Visual Dialogs

Hansjörg Mixdorff [1], *Angelika Hönemann*[2], *Jeesun Kim*[3], *Chris Davis*[3]

[1] Department of Computer Science and Media, Beuth University Berlin, Germany
[2]University of Bielefeld, CITEC, Germany,
[3] MARCS Institute, University of Western Sydney, Australia

mixdorff@beuth-hochschule.de, ahoenemann@techfak.uni-bielefeld.de,
[J.Kim:chris.davis]@uws.edu.au

Abstract

This paper presents an experiment in which we examined whether German and Australian English perceivers were able to predict imminent turn-switching in Australian English auditory-visual dialogs. Subjects were presented excerpts of one and four second duration either preceding a switch or taken from inside a turn and had to decide which condition they saw. Stimuli were either A/V, video-only or audio-only. Results on the one second excerpts were close to random. In general we found a preference for non-switching. Australian subjects outperformed the German subjects in the audio-only condition, but outcomes were almost equal on the A/V stimuli. Analysis regarding the syntactic and prosodic properties of the stimuli showed that phrase-final statement as well as question intonation facilitated recognition presumably due to these acting as markers of turn-switch preparation; whereas incomplete sentences and non-terminal intonation were indicative of turn-internal excerpts. As to visual cues signaling a following switch results were rather varied. An open mouth on the part of the listener more often preceded switches than not.

Index Terms: auditory-visual prosody, dialog, turn-switching

1. Introduction

It is well established that seeing a talker (visual speech) influences auditory speech processing. Typically, research has focused on the perception of segmental information and has demonstrated that visual speech facilitates speech perception [1]. Indeed, the McGurk effect shows that information processing from the two senses is strongly connected and conflicting cues are resolved to form the most likely percept [2]. It has also been shown that the provision of visual speech can improve the perception of lexical tone in noise [3] and can facilitate a range of perceptual judgments [4].

Moreover, recent research we have conducted suggests that visual speech influences the perception of speech prosody in interesting but possibly complex ways [5] . This work was based upon a corpus of spontaneous Auditory-Visual A/V monologs that was collected and annotated in terms of both acoustic as well as the visual properties. In addition, motion capture data was recorded and evaluated for non-verbal gestures.

In the analysis of this corpus, which involved the alignment of acoustic landmarks such as accents and boundaries with visible non-speech movements, the question arose as to which way the anchoring of movements should be achieved. In an initial approach only movements that occurred during accented syllables or syllables preceding a boundary were taken into account. However, this left a number of movements unanchored, where, for instance, these were located in syllables neighboring accented syllables. One limitation of the corpus collected in [5] was that it only consisted of monologs that had been delivered to a (mute) listener. Plausibly, non-verbal gestures may play an important role in structuring dialogs, so we decided to collect a corpus of spontaneous dialogs in order to examine more closely how non-verbal gestures facilitate discourse and interact with prosodic cues (e.g., in negotiating turn switches). In [6] we examined the structural properties of the data collected from 22 pairs of Australian English speakers engaged in “the Cartoon Task” (see below) in terms of contributions of the two speakers, turn duration and switching. We found a considerable amount of variation in the prosodic and visual expressions presented by the speakers. Some seemed to converge acoustically during the conversation, but no consistent patterns became visible, especially as indicators of an imminent turn-switching. For this reason we decided to approach this issue by means of a perceptual experiment. We extracted short scenes from the dialogs which either preceded a turn-switch or came from within a turn of one of the speakers. The remainder of this paper is structured as follows: In Section 2 we introduce the Cartoon Task and the collected corpus. Section 3 presents the details of the perception experiment. Section 4 discusses results from the experiment and the post-hoc acoustic and visual analysis of the stimuli. Section 5 offers discussion and conclusions.

2. Auditory-Visual Corpus

The *Cartoon Task* [6] was inspired by the Video Task developed by Benno Peters [7] and involves the interlocutors in a discussion about specially edited diverging versions of an episode of a soap opera. The resulting dialogs are relatively natural and balanced regarding the contributions of the two talkers. This task, however, requires that interlocutors are familiar with the particular series and also know each other well. The idea of discussing conflicting video presentations is appealing; however we wanted the task to be more focused and generalizable, i.e., not requiring any previous knowledge of the material or familiarity with the topic. Furthermore, since we ultimately plan to apply the same paradigm in different language and cultural environments, we selected an animated cartoon film of approximately eight minutes that had no dialog. Twenty-two pairs of participants (five of them male, 14 female and three mixed) were tested. Participants were recruited from the University of Western Sydney, aged between 17 and 53 and native speakers of Australian English. Participants were either students or university graduates. Most of the students participated for course credit, the remainder were paid. Two (approximately) five minute versions of the film were created in which the first and last scenes were common, but subsequent shots were present only in one or the other. In this way, the complete story was only recoverable when information from both versions was combined.

We informed participants that the experiment was about maintaining concentration and collaborating on a cognitive task. Participants were tested in pairs and were told that each person would view a different version of a short silent movie and that the versions were cut in such a way that they were going to see some scenes that their partner would not and vice versa. The cuts in the movie were made so that when a scene was missing the picture would cross-fade into the next scene and the missing scenes also recognizable by interruptions to the background music. We asked participants to memorize the sequence of events and the details of the scenes; they were told that subsequently they would be requested to interact with their partner in reconstructing the story. Specifically, participants were instructed that the story should be recovered cooperatively in chronological order and that they should avoid disclosing all the information they possessed at once, but rather piece together the sequence of scenes as the story develops.

For each participant of a dialog pair, 23 infra-red faces markers were applied in a standard configuration and three markers affixed to a head-worn rig (to track rigid head motion). Participants sat in a sound-treated room facing each other at a distance of about 1.5 m. Each was equipped with a head-worn microphone. Motion was captured with an eight camera Vicon motion capture system. Video was recorded with two Sony HDR-PJ200E HD video cameras manually (MPEG4-AVC/H.264 - 1920 x 1080/50i) (see Figure 1). The resulting two videos of each conversation were synchronized with the high quality audio from the motion capture system and joined in a single video that displayed both talkers alongside each other (see Figure 1) with the audio assigned to the left and right audio channel, respectively.

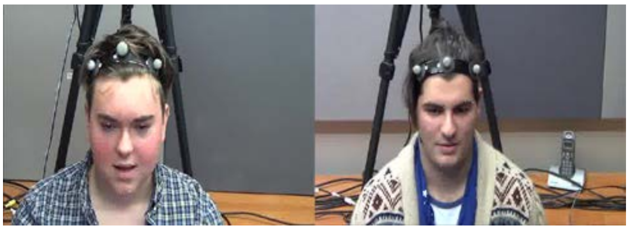

Figure 1: Combined videos of talkers A and B of Pair02.

3. The Perception Experiment

Based on a text level transcription of inter-pausal units performed in PRAAT [8] we identified turns and turn-switches. Using indicators such as total duration of the conversation and balancedness between speakers we selected ten pairs for extracting switch-preceding and non-switch preceding scenes. Once we had identified suitable locations we excised a set of excerpts of one and four seconds duration, respectively, before the prosodic phrase boundary preceding a pause that was either followed by a switch or not. This way we selected three switches and three non-switches each for every pair, yielding 10 x 6 x 2 (one and four seconds) = 120 stimuli. Of these 120 stimuli we randomly chose 60 that were also rendered in audio-only and video-only versions. The resulting 240 items were split into two sets of 120 each. These 120 were randomized individually for each participant. The experiment was programmed in JavaScript and run on desk top computers equipped with standard head-sets. 23 native German and 13 Australian English subjects took part in the experiment. 19 of them were students of Beuth University, 4 staff at University of Bielefeld and 13 students at University of Western Sydney, respectively, aged between 21 and 40.

In the auditory-visual and video-only conditions we showed the image of the two talkers alongside each other as depicted in Figure 1. Accordingly, the audio presented the voice of the left talker on the left head-set channel and the right talker on the right. Subjects required between 22 and 35 minutes for completing one set of stimuli. Some subjects commented that the one second A/V stimuli were too short to focus on who was talking, let alone identify whether a switch was imminent or not.

4. Results

First we have a look at recognition rates for the auditory-visual stimuli. The proportion correct for one and four second stimuli is listed in Table 1. As can be seen, with the one second stimuli, the performance for German and Australian subjects was similar, and approached chance level.

group	duration [s]	mean	s.d.
Australian	1	0.587	0.493
	4	0.680	0.467
German	1	0.560	0.497
	4	0.671	0.470

Table 1: Proportion correct for A/V stimuli.

There is only a moderate though significant correlation of 0.352 (Pearson's r, $p < 0.01$) between the responses and the binary non-switch/switch (0/1) property of the stimuli. If we look at the mean response for the two groups and types of stimuli, 0 being non-switch and 1 switch, we see a preference for non-switch, especially in the one second case (Table 2). As stated above, switch and non-switch stimuli were balanced in the sets. This indicates that when in doubt subjects opted for not having perceived a turn ending.

group	duration[s]	mean	s.d.
Australian	1	0.199	0.400
	4	0.428	0.495
German	1	0.266	0.442
	4	0.419	0.494

Table 2: Means and s.d. of responses for the two groups and A/V stimulus types.

Now we look at the results for switch and non-switch stimuli separately only considering the four second A/V stimuli (Table 3). Once again both groups performed similarly, and the preference for "non-switch" is obvious. Table 4 lists the proportions correct for the three types of 4 second stimuli. Here we only consider stimuli that had been presented in all three modalities. It is clear that subjects performed at chance level when presented with video-only stimuli. What is surprising is that, as shown before, the Australians performed similar to the German subjects on A/V stimuli. However, in contrast to the German subjects they do not exhibit an auditory-visual gain, that is, the Germans performed more poorly on audio-only and are almost at par with the Australians for the A/V stimuli.

switch	group	mean	s.d.
no	Australian	0.753	0.433
	German	0.751	0.433
yes	Australian	0.608	0.489
	German	0.590	0.493

Table 3: Proportion of correctly identified four second A/V stimuli.

mode	group	mean	s.d.
audio-only	Australian	0.612	0.488
	German	0.522	0.500
A/V	Australian	0.608	0.489
	German	0.620	0.486
video-only	Australian	0.509	0.501
	German	0.527	0.500

Table 4: Proportion of correctly identified four second stimuli depending on the modality.

Independent samples Mann Whitney-U-Test shows significant improvement ($p < 0.001$) of the Germans' performance from audio-only to A/V modality, whereas it is not significant ($p < 0.083$) for the Australians. In the A/V modality the responses of Germans subjects are strongly correlated with those of the Australians (Pearson's r=0.691, $p < 0.01$).

mode	group	mean	s.d.
audio-only	Australian	0.733	0.444
	German	0.554	0.499
A/V	Australian	0.759	0.430
	German	0.759	0.430
video-only	Australian	0.534	0.501
	German	0.580	0.496

Table 5: Proportion of correctly identified four second stimuli depending on the modality, "high-performers".

If we consider the individual performance of our subjects on the 4 second A/V stimuli we find that the proportion correct ranges between approximately chance-level at 46.4% to as much as 84.4%, regardless of the language group. Table 5 presents the results for the seven best performing subjects in each group, and the pattern shown confirms the observations regarding the Germans' A/V gain stated above.

Now we turn to the influence of the pair of talkers shown in the stimuli on the correctness of subjects' decisions, as we observed that some pairs are visually more explicit in the way they structure their discourse than others.

pair	mode	mean	s.d.
01	audio-only	0.696	0.465
	A/V	0.630	0.488
02	audio-only	0.379	0.490
	A/V	0.500	0.504
03	audio-only	0.500	0.508
	A/V	0.588	0.500
04	audio-only	0.400	0.497
	A/V	0.571	0.502
07	audio-only	0.655	0.480
	A/V	0.655	0.480
13	audio-only	0.471	0.507
	A/V	0.647	0.485
17	audio-only	0.429	0.502
	A/V	0.600	0.497
18	audio-only	0.413	0.498
	A/V	0.674	0.474
19	audio-only	0.652	0.487
	A/V	0.652	0.487
20	audio-only	0.617	0.491
	A/V	0.702	0.462

Table 6: Proportion of correctly identified four second stimuli depending on the pair of talkers.

Table 6 lists proportions correct for German subjects on 4 second stimuli depending on the pair and modality. As can be seen, not all pairs exhibit a similar A/V gain. Pair01 even has poorer recognition rates on A/V stimuli.

We now discuss our results of analysis regarding the acoustic and visual properties of the stimuli and how they relate to the underlying stimulus type, either preceding a turn-switch or non-switch. As regards the acoustic channel we labeled the following properties on the four second stimuli:

- prosodic realization of the stimulus end, either being declarative (statement), interrogative (question) or non-terminal (incomplete)
- syntactic property of the stimulus-final phrase as either being complete or incomplete
- the two talkers' speech activity (monolog of the turn-holding speaker throughout or not)
- back-channeling ('um', 'yeah', audible breathing etc.) by the listening partner

As can be expected, 20 of the 30 turn-final stimuli either exhibit declarative (16) or interrogative (4) prosody. In contrast, 24 of the non-turn-final 30 stimuli have a non-terminal ending. 27 of the turn-final stimuli end with syntactically intact phrases as opposed to only 14 of the non-turn-final ones. These results partially explain why the Australian subjects outperformed the Germans on audio-only stimuli as they could better draw on the acoustic cues. This idea is confirmed, for instance, by the observation that the

Australian listeners correctly identified 74.0% of switch-preceding audio-only stimuli with declarative prosody whereas that figure drops to 37.8% when the ending is non-terminal. Non-switches with non-terminal prosody were correctly identified 87.3% of the time. When we look at the presence of back-channeling we find that the majority of stimuli does not exhibit audible reactions by the listener, however, 11 switch-preceding stimuli contained back-channeling as opposed to 5 non-switch-preceding ones, some of it only being audible breathing. These figures are too low to draw conclusions, but may suggest increasing activity on the part of the listener when a turn-switch is imminent. As can be inferred from the relatively small number of listener reactions, most of the stimuli contain a true monolog by the turn-holding talker - which of course was the criterion for selecting the stimulus utterances.

With respect to the visual channel, a careful viewing of the stimuli did not reveal consistent cues associated with either switch or non-switch. We therefore decided to label the following stimulus properties for both the talker and the listener:

- presence/absence of eye brow movements
- presence/absence of body movements
- presence/absence of head movements
- gaze (at partner or not) at stimulus end
- mouth open at stimulus end or not

Evaluation of labels showed relatively few consistent behaviors. In 19 of the 30 non-switch stimuli, the talker exhibited head motion as opposed to only 8 instances in the turn-final stimuli. On the part of the listener, 11 turn-final stimuli ended with his/her mouth open (see left talker in Figure 1). Only two such instances were found in the non-final stimuli.

5. Discussion and Conclusions

In the current paper we examined whether German and Australian English subjects were able to distinguish between scenes preceding a switch of turns in a discourse and scenes taken from within a turn. To this effect we presented our subjects with one and four second stimuli of audio-only, video-only and A/V modality taken from an auditory-visual task involving two talkers. The data had been recorded at UWS Australia, the language being Australian English.

Performance of the subjects was close to chance level on the one second stimuli, as well as on the video-only ones. Therefore most of the ensuing analysis concentrated on the four second stimuli in audio-only and A/V modality. The Australian subjects performed similarly on the audio-only and A/V stimuli, whereas the Germans showed poorer performance when only the acoustic channel was present. However, they matched the Australians' recognition rates with the A/V data. This suggests that the German subjects drew on the visual information more than their Australian counterparts who presumably may have concentrated on the auditory channel due to their proficiency in the spoken language. This assumption is supported by the observation that most stimuli contained acoustically accessible cues such as prosodic or syntactic information. Complete syntactic phrases marked by a declarative or interrogative sentence mode signal that intervention by the listener is possible or even requested. In contrast, incomplete phrases with non-terminal prosody indicate that the talker is likely to continue, further elaborating on his point.

It is nonetheless surprising that the Australian listeners did not improve on their performance when visual cues were provided. This may indicate a certain threshold – possibly due to the inconclusiveness of some of the stimuli – beyond which the recognition rate could not rise. As mentioned earlier, the best performing subjects reached a score of 84.4% correct. The issue of precisely which visual cues the German subjects exploited will need to be evaluated in more detail. As we have shown, the A/V gain is not equal across talker pairs. This suggests that some talkers show more visually perceptable activity than others. Furthermore, a switch involves that the listener will start talking and therefore an open mouth appears to be a logical indicator for that to happen. As the stimuli were edited with respect to the acoustic channel, they often seem to have captured the (visual) preparatory phase of the ensuing turn.

In conclusion it must be stated that the task of classifying short video scenes regarding their position in a dialog is not a trivial one. Although full acoustic and visual information was offered, the viewing condition with both talkers side by side was counterintuitive as normally the talkers would be facing each other. It is, however, similar to situations we see on the TV when live-casts of several talkers are presented simultaneously. Nonetheless, it might have been more natural to present either the talker or the listener. We opted against this possibility as it would have doubled the number of stimuli. In future work, however, we will examine whether a single-talker display might be more effective, as the perceivers would be able to focus on only one face. Furthermore we will test how delexicalized or monotonized speech influences the judgments.

6. Acknowledgements

This work was supported by Deutsche Forschungsgemeinschaft international research grant Mi625/24 funding a stay by Mixdorff and Hönemann at MARCS Institute and a visit by Jeesun Kim at Beuth University Berlin. Special thanks go to Niels Garve, Denis Martin, and Nicolas Carosio, all master students at Beuth University, who edited the stimuli and programmed the experiment software.

7. References

[1] Sumby, W. H. and Pollack, I.,"Visual contribution to speech intelligibility in noise", JASA, 26, 212-215, 1954.

[2] McGurk, H. and MacDonald, J., "Hearing Lips and seeing voices", Nature, 264, 746-748, 1976.

[3] Mixdorff, H., Charnvivit, P. and Burnham, D., "Auditory-Visual Perception of Syllabic Tones in Thai", Proceedings of AVSP 2005, pp. 3 - 8, Parksville, Canada, 2005.

[4] Davis, C., & Kim, J. (2004). Audio-visual interactions with intact clearly audible speech. The Quarterly Journal of Experimental Psychology Section A, 57(6), 1103-1121.

[5] Hönemann, A., Mixdorff, H. and Fagel, S., "A preliminary analysis of prosodic features for a predictive model of facial movements in speech visualization", Proceedings of Nordic Prosody 2012, Tartu, Estonia, 2012.

[6] Mixdorff, H., Hönemann, A., Zelic, G., Kim, J., Davis, C. "The Cartoon Task – Exploring Auditory-Visual Prosody in Dialogs", Speech Prosody 2014, Dublin, Irland.

[7] Kohler, K. J., Peters, B. and Scheffers, M. (Eds.), "The Kiel Corpus of Spontaneous Speech IV, German: Video Task Scenario (Kiel-DVD1)", Kiel: IPDS, Christian-Albrechts-University, 2006.

[8] Boersma, P., "Praat, a system for doing phonetics by computer", Glot International, 5, 341-345, 2001.