The Cartoon Task – Exploring Auditory-Visual Prosody in Dialogs

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

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

[mixdorff:ahoenemann]@beuth-hochschule.de, [G.Zelic J.Kim:chris.davis]@uws.edu.au

Abstract

This paper introduces and analyses a collaborative task for eliciting auditory-visual dialogs based on the viewing of two versions of the same cartoon film. The original film was edited and cut in such a way that the story must be reconstructed by joining information from two incomplete versions which however share between them all the scenes in a consecutive fashion. Our intention is to elicit a relatively balanced dialog between the two participants throughout the conversation as they are piecing together the story from the beginning to the end. The current paper describes the production of the auditory-visual corpus using audio, video and motion capturing of 22 pairs of Australian English speaking participants, and presents first results regarding turn-distribution and raw prosodic features. Our analysis shows that the task is indeed relatively balanced between talkers though this does not apply equally to all pairs. Analysis of raw prosodic features does not suggest convergence throughout the conversation, but replicates, for instance earlier findings of similarity between partners as compared to others.

Index Terms: auditory-visual prosody, dialog, turn-taking, F0, intensity, entrainment

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 [1]. It has also been shown that the provision of visual speech can improve the perception of lexical tone in noise [3]. Moreover, recent research we have conducted suggests that visual speech influences the perception of speech prosody in interesting but possibly complex ways [4]. 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, a 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. In order to determine how the alignment of acoustic and visual cues reinforce the perceived prominence of the same underlying syllable(s), and when separate events of prominence are perceived, a perceptual rating experiment was designed in which the distance between auditory and visual cues for prominence was systematically varied [5]. The results of this work were in good agreement with a separate production study that examined the timing of head and eyebrow movement with respect to the expression of corrective focus [6].

At this stage, however, it is unclear how the results of the above controlled experimental studies applies to spoken dialogus, since a limitation of corpus collected in [4] 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 dialogues, 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 exchanges).

In the current study we examine this corpus with respect to the balancedness of speaker contributions. As a first application of the data we explore the effect of entrainment, the phenomenon that talkers engaged in a dialog adjust their speech to one another, e.g., such as synchronizing (turn-by-turn coordination between interlocutors), or where speech properties become more alike, that is, the talkers attain convergence [7].

The remainder of this paper is structured as follows: In Section 2 we introduce the cartoon task and the collected corpus. Section 3 presents statistical results based on the structures of the resulting dialogs. Section 4 discusses analyses and preliminary results regarding the prosodic entrainment between the participants in the dialog, as reflected by their *F0* and intensity contours, as well as their voice quality. Section 5 offers discussion and conclusions.

2. Experiment Design and Corpus

A large number of different paradigms exist for eliciting spontaneous dialog data. These paradigms range from completely unrestricted designs, in which at most only a general topic is given, to guided exchanges based on structured task solving. Some of the authors of this paper have applied the well-known Map Task [8] in their prosodic studies [9]. Although this task has been thoroughly studied and documented, its nature produces relatively unbalanced dialogs, as the Giver usually supplies most of the information for guiding the Follower to the desired location and the Follower's reactions often consist of one-word acknowledgments such as "yeah", "alright". In contrast, the Video Task developed by Benno Peters [10] 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.

2.1 *Participants*. 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 and knew each other previously. Most of the students participated for course credit, the remainder were paid.

2.2 *Materials*. 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.

2.3 *Procedure*. 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 DPA 4066-B head-worn microphone. In order for the facial markers not to be obscured the participants were asked not to raise their hands to their faces if possible.

After calibrating and adjusting the Vicon motion capture system (Lake Forest, CA) which consisted of 8 cameras (4 MX40; 4 MXF40), participants were provided with laptops and head phones for viewing the videos. After participants had finished viewing the video, we started two Sony HDR-PJ200E HD video cameras manually (MPEG4-AVC/H.264 - 1920 x 1080/50i) to have a visual record of each participant (see Figure 1). Following this, the motion capture system was started, capturing audio at 45kHz/16bit and marker motion at a frame rate of 100Hz and the participants were given a signal to begin. During the dialog no instruction were given to the participants. The recording was halted when the participants had decided that they had recovered the story as well as possible.

3. Analysis of Temporal Characteristics

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 along-side each other (see Figure 1). Then we performed text level transcription of inter-pausal units on the audio and also annotated non-verbal gestures such as audible breathing, smacks and laughter using the *Praat TextGrid* editor [11]. Based on these transcriptions we performed an analysis of talkers' contributions to the dialog in order to investigate whether the task was balanced.

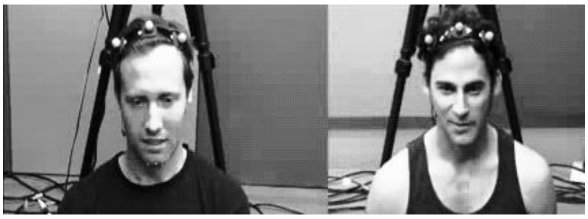

Figure 1: *Combined videos of talkers A and B of Pair01.*

Table 1 provides information on the resulting 22 dialogs, including the total durations, each participant's percentage of contributions, as well as the percentage of overlaps and silent pauses. Figure 2 displays sample graphic representations of turns along the time axis for a duration of four minutes. In each panel the black areas indicate activity of talker A and the grey areas indicate activity of talker B. As can be seen, Pairs 11 and 17 are balanced with regard to overall contributions by talkers A and B. However, the pairs differ greatly with respect to the distribution of turns. In Pair 11, both talkers produce longer stretches of speech and have fewer turn exchanges, in contrast to the talkers in Pair 17. This indicates that talkers apply different strategies for reconstructing the film. In Pair 11 talker A begins the dialog and talks about several scenes of his version, and only after that talker B presents his observations. The entire dialog continues in this way. Talkers in Pair 17 reconstruct the film more collaboratively by providing shorter pieces of information consecutively and ask each other for missing facts. They interrupt one another more frequently in order to take turns. This is the reason why Pair 11 exhibits only 7% of overlaps, but Pair 17 15%, as can be seen in Table 1. These two examples are representative for most others of the 22 dialogues. Both strategies to reconstruct the film appear to be successful technically, but Pair 17 obviously shared a more vivid exchange and followed our instructions better than Pair 11, hence providing more instances of turn exchanges that we wish to study. We checked whether the version of the video influenced the percentage of talkers' contributions. On average talker A speaks for 48%, and talker B for 41% of the total dialog time. Paired-samples T-test suggests that this tendency is small, but significant (T=2.239, df=21, $p < 0.036$).

4. Acoustic Analysis

For the subsequent analysis of prosodic features we selected ten pairs (nos. 1, 2, 3, 7, 11, 12, 17, 18, 20, and 22) where the contribution of the two talkers was relatively balanced and where the minimum discourse duration was at least four minutes.

Due to the close proximity between the two talkers during the recording there was audible cross-talk in each of the audio channels, the channel separation being approximately 15-20 dB. By applying audio source separation [12] we yielded a gain of 6-8dB without audible deterioration of the speech signal.

Table 1: *Overview of the 22 dialogs with total durations, percentage talking time of talkers A and B, percentage common pauses and overlap between A and B.*

#	total dur. [s]	% A	% B	% common pause	% overlap	#	total dur. [s]	% A	% B	% common pause	% overlap
01	524	43	37	30	10	12	349	52	40	19	11
02	256	39	37	33	9	13	273	68	36	9	13
03	349	44	40	27	11	14	312	62	33	21	16
04	293	48	28	32	9	15	296	52	38	22	12
05	274	28	64	17	9	16	109	37	49	17	3
06	113	34	41	37	12	17	403	54	50	11	15
07	288	51	45	19	14	18	428	56	41	11	8
08	101	45	24	37	6	19	212	48	50	18	17
09	265	54	25	26	6	20	264	44	52	17	13
10	290	43	38	31	11	21	181	63	35	17	15
11	275	42	48	17	7	22	287	57	43	16	15

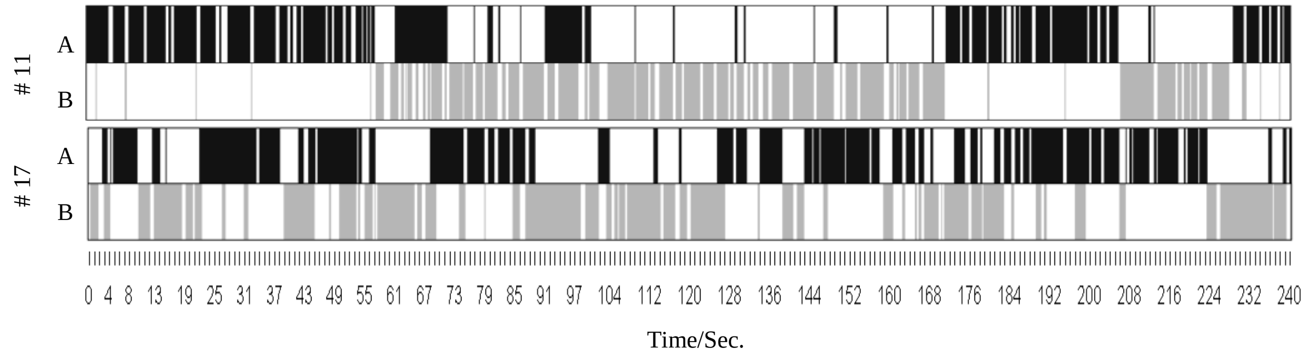

Figure 2: *Graphic representations of turns for two selected conversations displayed for chunks of four minutes. In each panel the black areas indicate activity of talker A and the grey areas indicate activity of talker B.*

We then extracted *F0* contours at a step of 10ms employing the *Praat* standard algorithm [11] with different *F0* floors and ceilings for male (50-300Hz) and female participants (130-400Hz). Along with the *F0* values the *Praat PitchObject* contains information on frame intensity as well as periodicity, a measure comparable to the harmonics-to-noise ratio. For each of the features *F0*, intensity and periodicity we calculated z-scores with respect to the male and female mean and standard deviations, respectively.

In principle our analysis follows the approach presented in [7]: in order to examine the prosodic entrainment between the two talkers in each conversation we calculated means, standard deviation as well as minimum and maximum values of the resulting feature z-scores for chunks of constant length in the conversations (since we do not yet dispose of a detailed transcription of inter-pausal units as well as annotation of turn exchanges).

We then performed two types of analysis: (1) Correlation analysis between the sequences of chunk-wise features for the entire conversation; (2) Statistical analysis of absolute differences between chunk-wise features depending on the talker, the pair, the distance between chunks, as well as the start time of the chunk with respect to the conversation.

After experimenting with several chunk sizes we employed durations of 20s for our subsequent tests. In order to ensure that chunk parameter sets contained averaged values from a sufficient number of speech frames, we required a chunk to contain at least 6s of speech by the talker examined, a speech frame being defined by the intensity reaching a fixed threshold.

On the conversation level, as a test for proximity, we calculated the correlations between sequences of chunk parameters by partners as well as non-partners. Since conversations varied in length, the number N of chunks employed for each analysis varied as well. Results are displayed in Table 2.

Table 2: *Conversation-wise inter-partner correlations (Pearson's r) of mean intensity and mean F0 for selected pairs.*

pair	N	r(mean int.)	p	r(mean F0)	P
01	24	.70	.001	.37	.072
02	10	.72	.020	-.12	n.s.
03	14	.91	.001	.53	.053
07	8	.11	n.s.	.86	.007
11	4	-.97	.035	.93	.067
12	10	-.10	n.s.	.36	n.s.
17	18	.31	n.s.	.39	n.s.
18	15	.39	n.s.	-.02	n.s.
20	8	.34	n.s.	.47	n.s.
22	12	.18	n.s.	.79	.003

Of all features only mean intensity and mean *F0* yielded inter-partner correlations that were significant or approached significance for some of the pairs. Results for pair 11 may be unreliable due to the small number of chunks in which both partners have a sufficiently high number of speech frames.

In our analysis of chunk-wise parameter differences we first calculated means and standard deviations of feature differences between chunks of the same talker (*self*) as compared to those by others (*other*, see Table 3). As expected, talkers were much more similar to themselves than to others.

We then performed intra-talker correlation analysis of chunk-wise feature differences as a function of the distance between the chunks compared. Only mean intensity (Pearson's r =.11, p <.001), intensity s.d. (r = 0.10, p < 0.005) and mean periodicity (r = 0.09, p < 0.02) indicated a weak tendency of the talker to be more dissimilar to him/herself between chunks in discourses that were spaced further apart.

Table 3: *Feature difference means and standard deviations self vs. other.*

	F0 mean	F0 sd	F0 max	int. mean	int. sd	int. max	per. mean	per. sd	per. max
self mean	.24	.22	1.67	.24	.22	1.04	.17	.09	.01
self s.d.	.24	.18	1.25	.22	.21	.92	.14	.08	.08
other mean	.55	.28	2.08	.35	.31	1.50	.29	.16	.18
other s.d.	.42	.22	1.46	.29	.27	1.07	.23	.14	.17

Turning to the relationship between talkers who were engaged in the same conversation (*partner*) as opposed to those in others (*other*), we conducted T-Tests on chunk differences. The results shown in Table 4 indicate that for most features the differences between talkers in the same conversation (*partner*) were smaller compared with talkers from a different conversation. For mean *F0* the difference between *partner* and *other* was significant as well, though the feature differences proper were larger between partners of the same pair.

An intra-pair correlation analysis of chunk-wise inter-talker differences was performed to see whether chunks spaced further apart were more dissimilar. However, we only found a rather weak dependency of mean *F0* on the distance between chunks (Pearson's r = 0.13, p < 0.002).

If we compare intra-pair chunk-wise parameter differences as a function of the onset times of the chunks by only including pairs of chunks occurring at the same time or neighboring one another, mean *F0* (r = 0.28, p < 0.001), intensity max (r = -0.20, p < 0.004) and periodicity s.d. (r = -0.22, p < 0.002) were weakly correlated with the onset times of the chunks in the conversation.

Table 4: *T-tests partner vs. other differences*

Feature	t	df	p-value	Sig.
intensity max	-2.8	7273	<0.001	*
intensity mean	-3.1	7273	0.002	
intensity s.d.	-3.2	7273	0.001	*
F0 max	-5.3	7273	<0.001	*
F0 mean	10.5	7273	<0.001	*
F0 s.d.	-2.7	7273	<0.001	*
periodicity max	-20.4	7273	0.005	
periodicity mean	-0.2	7273	N.S.	
periodicity s.d.	-2.5	7273	0.012	*

As a test of whether or not talkers in the same pair converged during the conversation we examined chunk differences calculated for chunks located in minutes 1 and 2 of the discourse with those in minutes 3 and 4, as only a few of the conversations were considerably longer than four minutes. Mann-Whitney independent sample U-Test suggests differences for intensity mean (p < 0.039) and intensity max (p < 0.017), however, the tendency was for talkers to become more dissimilar with respect to these features later in the discourse.

5. Discussion and Conclusions

This paper presented the first results from an auditory-visual corpus of spontaneous dialogs based on a collaborative task centered on the reconstruction of a cartoon film. Based on transcriptions of inter-pausal units and the inspection of graphical representations of discourse structures we found that conversations overall are relatively balanced between talkers, although pairs differed with respect to the total duration of the discourse as well as turn durations and the amount of overlap.

For a subgroup of relatively well-balanced pairs we examined the prosodic features *F0*, intensity and periodicity with respect to entrainment, the adaptation that can occur between talkers engaged in a conversation. We calculated chunk-wise means, standard deviations as well as min and max values of feature z-scores and examined the relationships between these features for chunks of 20s length. With respect to the whole discourse intensity means exhibited the highest correlations between partners, followed by *F0* max, however this was the case only in some of the pairs. This might reflect individual differences in discourse strategy between pairs. For example, in some dialogues, one partner took the lead and presented most of the information s/he had before granting a turn exchange. In these cases adjustment by the partner may be more difficult than in pairs where the information was delivered in balanced turns.

We investigated chunk-wise feature differences between talkers and themselves, their partners and talkers with whom they had not conversed. With respect to a number of features, especially intensity and *F0*, talkers were more similar to their partners than to other talkers. The similarity seemed to decrease with the distance between chunks in time, though the dependency was relatively weak. We did not find evidence of talkers converging during a conversation though this might simply be due to the short durations of most dialogs. It rather seemed that talkers diverged with respect to intensity, for instance. We believe that it will be necessary to perform a detailed annotation of turns and turn exchanges to better pin-point possible places of stronger coordination. We also require word, syllable and phone-based segmentations in order to test for the entrainment of duration information. In addition, future work will involve annotations of non-verbal facial or head movements followed by the analysis and modeling of the motion capture data.

6. Acknowledgements

We would like to thank Simone Simonetti at MARCS Institute for organizing participants and checking transcriptions and Yuanfu Liao of NTUT Taiwan for applying channel separation to the audio recordings. This work was supported by Deutsche Forschungsgemeinschaft international research grant Mi625/24 funding a stay by Mixdorff and Hönemann at MARCS Institute.

7. References

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

[2] McGurk, H., & MacDonald, J., “Hearing Lips and seeing voices”, in: Nature, Volume 264, pp. 746-748, 1976.

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

[4] 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.

[5] Mixdorff, H., Hönemann, A. and Fagel, S., “Integration of Acoustic and Visual Cues in Prominence Perception”, Proceedings of AVSP 2013. Annecy, France, 2013.

[6] Kim, J., Cvejic, E. and Davis, C.,“Tracking eyebrows and head gestures associated with spoken prosody. Speech Communication, 57, 317–330, 2014.

[7] Levitan, R. and Hirschberg, J., “Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions”, Proceedings of Interspeech 2011. Florence, Italy, 2011.

[8] Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G.M., Garrod, S., Isard, S. D., Kowtko, J. C., McAllister, J., Miller, J., Sotillo, C. F., Thompson, H. S. & Weinert, R., “The HCRC Map Task Corpus. In: Language and Speech 34, pp. 351-366, 1991.

[9] Mixdorff, H., Pech, U., Davis, C. and Kim, J., “Map Task Dialogs in Noise - a Paradigm for Examining Lombard speech”. Proceedings of ICPHS07, Saarbrücken, Germany, 2007.

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

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

[12] Ozerov, O. and Févotte, C. “Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation”, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 18, NO. 3, MARCH 2010.