A Verbal Interaction Measure Using Acoustic Signal Correlation for Dyadic Cooperation Support

Alexander Neumann and Thomas Hermann

Abstract We introduce a method for detecting whether two users are engaged in focused interaction using a windowed correlation measure on their acoustic signals, assuming that a continued exchange of verbal turns contributes to anticorrelation of acoustic activity. We tested our method with manually annotated transitions between focused and unfocused interaction stemming from experiments on AR-based cooperation within a research project on alignment in communication. The results show that a high degree and extended duration of speech activity anticorrelation reliably indicates focused interaction, and might thus be a valuable asset for situation-aware technical systems.

Key words: situation awareness; collaboration; speech activity; data mining; multiscale analysis; correlation

1 Introduction

Recent developments on technical interactive systems do not only focus on user interfaces that are easy to use but also take the actual usage context into account. Features like ambient light or GPS location information are already used to change the behavior of mobile phones or smart environments, allowing to create situation awareness and adapt the system to the changing environment. Verbal utterances are commonly used either in a rudimentary way to detect general ambient noise or in a complex way which involves speech recognition and semantic parsing. Regard-

Alexander Neumann
Bielefeld University, Universitaetsstrasse 25, 33615 Bielefeld, Germany
e-mail: alneuman@cit-ec.uni-bielefeld.de

Thomas Hermann
Bielefeld University, Universitaetsstrasse 25, 33615 Bielefeld, Germany
e-mail: thermann@techfak.uni-bielefeld.de

ing conversation, using speech only for noise detection ignores its vital role in joint activities [2], while speech recognition often does not fulfill accuracy or speed requirements for reliable information gathering in such a context. It also demands powerful hardware which can involve more than one recording device [7].

We propose a simple and lightweight speech activity correlation approach to reveal verbal dialogue communication patterns which can be used to increase situation awareness for static and mobile cognitive interaction technology. These developments come from the *Augmented Reality based Interception Interface* (ARbInI) which we developed as a system to investigate communication phenomena such as alignment, joint attention and co-orientation in human-human interaction. *ARbInI* was used to collect data in our latest study, a cooperative interaction study where participants had to collaboratively plan fictional building activities and negotiate possible solutions. Besides video and tracking data, the recorded multimodal corpus also includes sound signals from headset microphones that our participants had worn during the experiments.

Exploratory data mining revealed interesting speech activity patterns in these data which we further investigated. Based on 10 dyads from our corpus we developed a correlation measure which depends on noise threshold Θ, silence duration d_p and correlation window size ω, and we tested the algorithm performance against manual annotations of the same data. In the following we will give a brief introduction to our study and the collected data corpus. After that, we will introduce the algorithm and its evaluation and furthermore also show how the algorithm's parameters can be determined from the data. However, we propose that d_p and ω do not have to be adapted to fit varying scenarios.

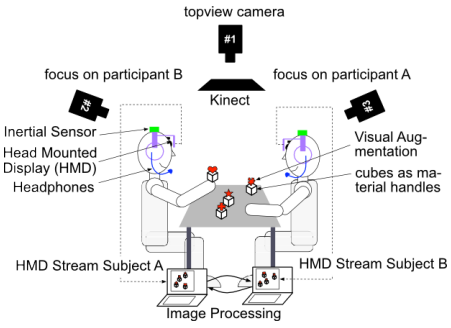

Fig. 1 ARbInI consists of static components such as three DV cameras, a Microsoft Kinect and two to three workstations. Each participant also wears a head-mounted display, a microphone headset and a BRIX motion sensor to measure head movement at high temporal resolution.

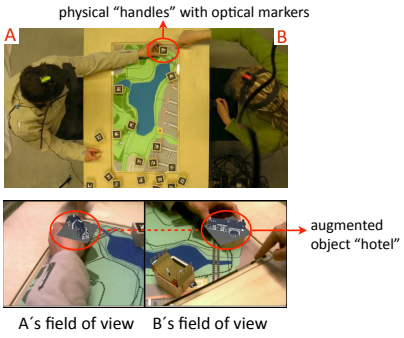

Fig. 2 In the ongoing study our participants collaborate to recreate a local lake and its surroundings. *ARbInI* monitors their actions. The markers on top of the wooden cubes are augmented with models representing concepts for possible projects (e.g. *hotel* or *skater park*).

2 Alignment in AR-based Cooperation

The Collaborative Research Center 673 *Alignment in Communication*[1] investigates the role of alignment and other communication patterns for successful communication. In the subproject C5 *Alignment in AR-based collaboration* we use Augmented Reality (AR) as a technology for communication research which provides new features and methods for this discipline.

Within this context the *Augmented Reality based Interception Interface* (ARbInI) was developed and tested as a monitoring and assistance system in everyday dialogue scenarios [4]. The system allows a direct access to the audiovisual communication channels to monitor and alter information perceived by the users. Combined with other non-verbal communication cues such as gestures, posture and gaze direction these data form a complex multimodal data corpus.

2.1 ARbInI and Obersee II Scenario

Our system consists of several components which are either positioned around two chairs and a table or worn by the users. All components are shown in Figure 1. The sensors attached to the users contain motion sensing devices from the BRIX toolkit which was developed in our working group [10] and headset microphones to record audio signals. The core component is a video-see-through head-mounted display (HMD) equipped with two Firewire cameras and a display for each eye. Three HD digital video cameras surround the participants, two of them are placed diagonally behind each participant and the third right above the table where also a Microsoft Kinect[2] is located. All data streams can be accessed, stored and manipulated in real-time except for the HD videos which we only record for later analysis.

For the study we have designed a recreation planning scenario which takes place in the surroundings of a lake called Obersee in the city of Bielefeld.

Figure 2 shows the setup from the top with the sketch of the Obersee area in the middle of the table. An important part for our AR approach is the introduction of mediating objects which represent constructions for the participants to use for their planning. They are wooden cubes which are used as “physical handles” with *ARToolkitPlus* [8] markers attached on top. When the system detects a marker it augments the corresponding visual representation of a building or concept on top of the cube as depicted in Figure 2. This feature allows us to monitor, control and manipulate the visual information available to both users separately during the negotiation process at every moment during the experiment [3].

[1] www.sfb673.org

[2] www.xbox.com/en-US/kinect

Fig. 3 This is participant B's waveform of a 5 second dialogue sample. Participant A interrupts participant B to deny her suggestion instantly. The orange area of the waveform indicates parts louder than −15 dB which is a sufficient threshold choice here. The speech activity before the interrupt should be merged into a continuous activity but the pause must remain. A short silence duration like 200 ms leaves the activity fragmented; a long one like 1500 ms might close too many gaps.

3 Analysis

In the analysis process of the collected data, we investigated speech activity as a feature for measuring the degree of collaboration. We define *speech activity* as any verbal utterance which addresses the speaker's interlocutor with no regards to syntactic or semantic information.

We retrieved speech activity from the subjects' microphone recordings with a *sound finder* based on an audacity plugin by Jeremy R. Brown[3]. This approach reads 100 samples of a signal and detects the sample with the highest volume within this frame k and returns 1 if this sample is louder than Θ. The result was further compressed with a *sample & hold* interpolation to fit the 50 Hz sample rate of our data set.

$$\begin{aligned} sp[k] &= \begin{cases} 1 \text{ if } 10 \cdot \log_{10}(\max(s[i]^2)) > \Theta \\ 0 \text{ else} \end{cases} \\ s[i] \in [-1,1], \frac{i}{100} &\in [k,k+1], t = \frac{k}{441}\text{sec} \end{aligned} \tag{1}$$

In our case −15 dB has been proven to be a robust and reliable noise threshold which detected all verbal utterances articulated by the speaker without *false positives* like background noise, speech activity of the interlocutor or pure intrapersonal stimulation such as very quiet "hmm" sounds which did not fulfill communication purposes. Certainly, this threshold depends on our special case since used hardware and control parameters (e.g. microphone volume) vary between scenarios.

However, the feature so far leads to fragmented results. For instance even a single word like "friendship" could result in two chunks due to intonation and short pausing between syllables. Therefore, we applied an *erosion* method where gaps within a continued activity are bridged and fragments are merged into a continuous segment if the gap is shorter than a silence duration parameter d_p. We chose this duration with help of the multiscale correlation structure described in section 3.1. Our goal was to ignore small pauses (e.g. "well,... uhm... what about here") but to keep independent statements separated as depicted in Figure 3.

[3] audacity.sourceforge.net

3.1 Structure in Verbal Dyadic Interaction

To better understand the distribution of gaps and the effect of erosion on the stability of utterance lengths we introduced a multiscale analysis of acoustic segment statistics. Specifically we coupled the histogram of segment length as a function of the erosion length d_p. Figure 4 depicts the result using a log color mapping for frequency, and showing d_p on the y-axis for a participant. Interestingly there are vertical bars at certain segment lengths, corresponding to repeated occurrences of specific utterance durations which remain quite stable under variation of d_p and gets more visible when pooled data of 20 participants is used. This (visually) suggests a corridor of $d_p \in [250, 1500]$ ms in which stable statements are rarely affected by the erosion approach.

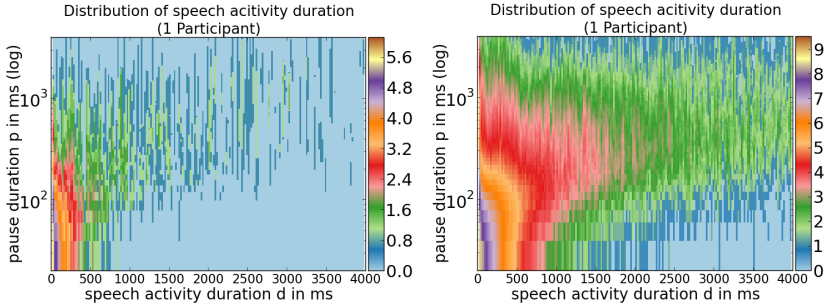

Fig. 4 The color maps show the distribution of speech activity durations on the x-axis and the chosen threshold d_p on the y-axis. The color is the logarithm of the amount of activities with a certain duration. The left plot shows the data of one participant which includes some vertical lines, for instance at 2000, 2500 or 3500 ms of speech activity duration. These lines indicate durations which are very consistent for d_p in range of 250 to 1500 ms. With 20 participants included (as seen on the right) this lines form a "corridor" within this range.

3.2 Correlation

The processed data of both participants are used to calculate a windowed correlation as function of sample time k using equation (2).

$$\mathrm{corr}_{xy}(k,\omega) = \frac{4}{\omega} \sum_{i=k-\omega}^{k} (sp_x[i] - 0.5) \cdot (sp_y[i] - 0.5) \tag{2}$$

We use a rectangular window function centered at $t = 0$. Local structure decreases with increased window size ω and stabilizes so that fast oscillations are filtered since the windowing operates similar to a low pass filter on the product feature $x \cdot y$. The correlation is computed on the *speech activity* feature introduced in section 3. Different from standard correlation, we shift the features so that silence is represented by $-\frac{1}{2}$ and speaking by $+\frac{1}{2}$. The motivation is that both joint silence and joint speaking should contribute in equal measure to positive correlation. For the correlation function to range between -1 and $+1$ the result is multiplied by 4.

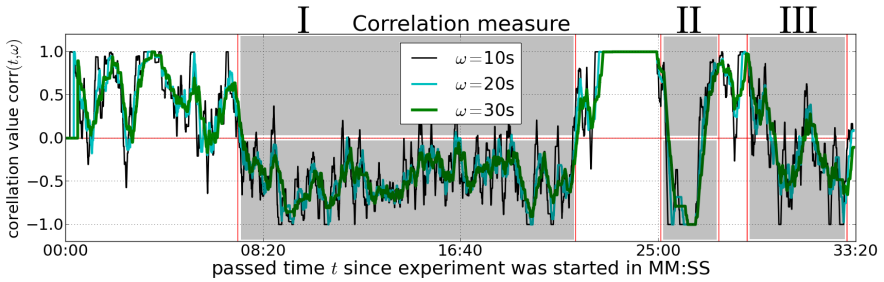

Fig. 5 The graph shows the correlation result of the participants' speech activity time series. The vertical red lines mark phase transitions which were annotated manually. The numbers mark the negotiation (I), presentation (II) and free phase (III) of the experiment where participants collaborated. The correlation changes during phase transitions where the focus shifts from the interlocutor to the experimenter or vice versa.

4 Evaluation

To evaluate the correlation results, we manually annotated phases in our video data which occurred in a certain order in our experiment. We started with an *introduction phase* where the setting and task was introduced by the experimenter and the participants mostly listened or talked to a person from the experiment team. In the *negotiation phase* the participants had to discuss and agree on solutions for the recreation planning task. The study personnel left the room during that phase. The negotiation phase ended when the participants rang a bell and was followed by a presentation of the final solution. Between negotiation and presentation was a small window where the experimenter asked some question and handed out a questionnaire. After the presentation, the staff left the room a second time for about 5 minutes which was called the *free phase* where the participants were left sitting on the table without the mediating objects to record pure conversation data[4].

In Figure 5 both results are shown together for one trial exemplary. The correlation graph's zero crossing happens shortly after the negotiation and the presentation phase started and stays below zero right until the end of the phase. In the free phase we observe more fluctuation which additionally differs for every trial.

We cross-validated our findings by correlating time series from different trials to verify the approach. This was only done for phase transitions since these are essential moments for conversation detection and the phases' durations varied across the trials. Figure 6 shows such a cross-validation for start and ending of the negotiation phase. Speech activity time series s_1 and s_2 belong to the same trial that was shown in Figure 5 and were checked against s_e from another one. The blue graphs depict the inter-trial correlation and show similar shapes as the green graph before and after the end of the negotiation phase. During the phase the graph passes zero several times and fluctuates within the range of about $-\frac{1}{2}$ and $+\frac{1}{2}$ in both cases (s_1, s_e) and (s_2, s_e). The gaps are a result of the differing length of the negotiation phases.

[4] The participants were told that some system calibration had to be done to finish the experiment

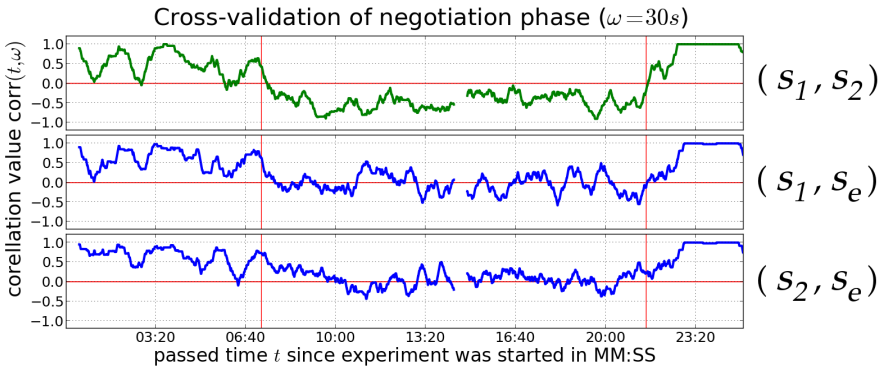

Fig. 6 The graphic shows the correlation for the time series s_1, s_2 from one trial and s_e from another trial. The vertical red lines mark the start and end of the negotiation phase which was annotated manually. The time series s_1 and s_2 (green graph) constantly anticorrelate after the phase transition until the end of the phase. Pre- and post-negotiation results of (s_1, s_e) and (s_2, s_e) look similar to the top graph, but steady anticorrelation cannot be observed during this phase. The gap at 13:40 is caused by different negotiation phase durations of the trials.

5 Discussion

Collaboration requires listening and a proper turn taking behavior where overlaps are accepted (in contrast to interrupts) and cause minor speech activity correlation. One person should speak at a time even though research has shown that there can be overlap towards the end of a turn or for backchannling (e.g. “yeah.. ahh”) depending on the social norm, context and the interlocutors’ relationship [5]. Weilhammer and Rabold found that average overlap related to the spoken language ranging from 150 to 330 ms for English, German and Japanese speakers [9].

The fact that cooperative speaking behavior anticorrelates is not surprising. But it is interesting how accurate this feature alone can determine if both participants cooperate. In our trials the probability of cooperative interaction was tightly coupled to the degree of the participants’ verbal anticorrelation and its duration. Values smaller than $-\frac{1}{2}$ were hardly reached by cross-correlated time series.

For more fractured conversations this approach has to be adapted since the turn-taking time (also called inter-speaker interval) depends on the task [1]. The similarity of the inter-trial correlation with the intra-trial correlation shown in Figure 6 indicates that during those periods all participants, disregarding the trial, were listening most of the time to the experimenter’s instructions which is supported by our qualitative analysis of the data. Joint silence is treated as uncooperative which is okay if both participants listen to the experimenter but it does not have to be true in all situations. We believe that this is one reason for fluctuation during the negotiation phase. Suppressing this behavior has to be done very carefully since in some cases the lack of verbal communication can be an indicator for recent problems in the problem solving process.

6 Conclusion

We have introduced and tested a new reliable signal-driven method for interaction focus detection from speech signal correlation. However, joint silence is treated as correlation and thus influences the current rating heavily. We propose a memory-based weight-decay feature to take the likeliness of a conversation between two (or more) interlocutors into account.

This approach may be useful to improve context awareness of future devices, a factor of increasing relevance in application development [6]. Importantly, this feature can be computed without any privacy-intrusion as no semantic features are accessed. Until full speech recognition-based interaction analsyis becomes available and cheap, our approach can support real-time situation detection.

As an interesting application beyond the scope of this paper we suggest the ubiquitous *Chatter Tracker* for parties, conferences or other social events: Every mobile phone running the application would collect speech activity to a server, which in turn computes pairwise correlations and composes for each interlocutor a summary of whom he has spoken with. Never forget to exchange contact information again as this could replace business cards.

Acknowledgements This work has partially been supported by the Collaborative Research Center (SFB) 673 Alignment in Communication and the Center of Excellence for Cognitive Interaction Technology (CITEC). Both are funded by the German Research Foundation (DFG).

References

1. Bull, M., Aylett, M.: An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. In: Proceedings of ICSLP (1998)
2. Clark, H.: Using language, vol. 4. Cambridge University Press Cambridge (1996)
3. Dierker, A., Mertes, C., Hermann, T., Hanheide, M., Sagerer, G.: Mediated attention with multimodal augmented reality. In: Proceedings of ICMI-MLMI, p. 245. ACM Press, New York, USA (2009)
4. Dierker, A., Pitsch, K., Hermann, T.: An augmented-reality-based scenario for the collaborative construction of an interactive museum. Tech. rep., Bielefeld University (2011)
5. Edelsky, C.: Who's got the floor. Language in society **10**(3), 383–421 (1981)
6. Grudin, J.: The Computer Reaches Out: The Historical Continuity of Interface Design. Proceedings of CHI '90 pp. 261–268 (1990)
7. Lecouteux, B., Vacher, M., Portet, F.: Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions. In: Proceedings of Interspeech, pp. 2273–2276 (2011)
8. Wagner, D., Schmalstieg, D.: Artoolkitplus for pose tracking on mobile devices. In: Proceedings of CVWW (2007)
9. Weilhammer, K., Rabold, S.: Durational aspects in turn taking. International Congresses of Phonetic Sciences (2003)
10. Zehe, S.: BRIX - An Easy-to-Use Modular Sensor and Actuator Prototyping Toolkit. In: Proceedings of SeNAmI 2012, pp. 823–828. Lugano, Switzerland (2012)