Micro-Timing of Backchannels in Human-Robot Interaction

Benjamin Inden
Artificial Intelligence Group
Bielefeld University
Universitätsstr. 25, 33615
Bielefeld, Germany
binden@techfak.uni-bielefeld.de

Zofia Malisz
Petra Wagner
Faculty of Linguistics and Literary Studies, Bielefeld University
{zofia.malisz, petra.wagner}@uni-bielefeld.de

Ipke Wachsmuth
Artificial Intelligence Group
Bielefeld University
ipke@techfak.uni-bielefeld.de

What is micro-timing of backchannels? Why should it be relevant to HRI?

Timing is important for all aspects of human-human and human-machine interaction. For example, collaborative action in various tasks becomes successful through well coordinated sequences of actions. Of course, dialogues, as a particular form of joint action, benefit from timing as well. It helps to avoid overlapping speech and long pauses, which makes the participants feel more positive about the conversation as well as the conversation partner [36, 7].

Our research has so far been conducted in the context of embodied conversational agents, where timing has been identified a crucial issue early-on for gesture [33], with applications also in HRI [31], and in the conversational agents domain. It more recently focused on backchanneling [11]. These investigations should also be relevant for making human-robot interaction more natural and pleasant.

There is already a vast body of literature on the importance of timing, and on strategies for timing gestures and verbal utterances. If timing is viewed as a single level phenomenon, then quality of interaction can be viewed as increasing more or less linearly with timing accuracy. However, we propose the existence of at least two distinct levels of timing that are sensitive to different influences, and are processed cognitively in different ways. We call the lowest level of timing, which we assume to be based on entrainment, *micro-timing* as compared to macro-timing, which takes place on coarser time scales and is likely to involve more conscious, strategic processing. We propose that to achieve a high quality of interaction, one should pay attention to timing on both levels.

Here we are particularly interested in one aspect of timing in dialogue: The timing of *backchannels*. Backchannels are brief signals produced by the listener in an asymmetric dialogue to signal attention and encourage the speaker to go on. Head nodding is one of the most unobtrusive forms of backchannels. Brief utterances like 'um', 'ok', or 'yeah' can serve as verbal backchannels, or for higher functions of feedback. Various studies have convincingly shown that conversational agents are viewed by humans as more natural and likable, and dialogues with them are more successful, if they produce backchannels, and use empirically derived strategies for timing of backchannels [16, 21, 28]. For macro-timing of backchannels, verbal backchannels are typically placed towards the end of the utterance and into pauses of the speaker. Often, these backchannel opportunity windows are initiated by pitch changes in the partner's speech [28]. Our own work aims to complement existing strategies by paying attention to micro-timing, thus achieving a higher level of success in human-machine interaction. Macro-timing provides windows of opportunity for backchannels, within which micro-timing is used to determine the actual time of production. Our working hypothesis for micro-timing is that prosodic events like vowel onsets, or vowel onsets in stressed syllables, can drive an oscillator that provides good candidate timings for the onset of a listener's backchannel.

Timing in Interaction, Workshop in Conjunction with HRI 2014, March 3, 2014, Bielefeld, Germany.

What is micro-timing based on?

Coordination can be achieved by predicting the partner's actions, including his or her turn. Theories of timing by prediction often rely on two mechanisms: Firstly, the partner's words and actions activate certain concepts by association (this is called priming, and takes place in the cortex) [27]. These activations also take place in the motor areas of the brain (an important point in theories about mirror neurons [30]), and allow predictions about what comes next. Secondly, once particular motor action concepts are activated, they can trigger simulations about how an action will occur, and how long it will take [5] (according to motor control theory, these simulations involve the cerebellum, and are improved by supervised learning [40]).

A different kind of mechanism can also be used for predicting the timing of quasiperiodic events. Oscillators produce periodic activations. If they are coupled to external periodic signals, many types of oscillators will adapt their phase and period to the external signal (a process that is called entrainment), thereby becoming suitable predictors. Some oscillator models can deal with rhythms (i.e., events with hierarchically structured regularities — like our postulated levels of timing in dialogue) because on the one hand, they can entrain to signals not just at a period ratio of 1:1, but also at more complex integer ratios [17], and on the other hand, oscillators entrained to periodic signals on different levels of the rhythmic hierarchy can be coupled, thereby influencing

each other and potentially achieving a higher overall prediction performance [24]. In addition, if a bank of oscillators tuned to different frequencies is present, they can entrain to different levels of the rhythm simultaneously, and deal with abrupt changes in frequencies rather well [18, 10]. Oscillator models have been shown to be able to model humans' competence to rhythmically entrain to musical signals. They even produce rhythms in music that do not directly correspond to any frequency found in the spectral analysis of the signals, but which are perceived by humans as well [18, 12]. An example how an oscillator bank reacts to vowel onsets in spontaneous speech can be seen in Fig. 1.

What makes us think that entrained oscillators could be a mechanism for micro-timing of backchannels? Entrainment phenomena have been detected on various levels of interspeaker coordination, e.g. in synchronous speech reaching very high temporal agreement [2], in the timing of overlapping speech [37], or in postural swing [29], but also for various other group interactions such as spontaneous rhythmic clapping [23]. Recently, evidence for the existence of inter-speaker entrainment in natural dialogues across various, typologically and rhythmically very different languages was provided [37, 38, 39]. More specifically, it was shown that overlapping utterances are not distributed randomly across an interlocutor's turn. Rather, they tend to be in phase with vocalic onsets or perceptual centers of the interlocutor's speech, but are least likely to be initiated around pitch accents. The link to perceptual centers and pitch accents at least hints at two levels of the prosodic hierarchy, namely the syllable, closely tied to the intervocalic interval, and the prosodic foot, often operationalized as the interval between accents.

When measuring brain activity, one can find a number of prominent frequencies in the range of typical speech units and important windows of psychoacoustic processing: the theta band (3-12 Hz) roughly corresponds to the typical duration of a syllable (100-300 ms), whereas delta band oscillations (0.5-3 Hz) correspond to typical lengths of prosodic and metrical units. It has also been shown that endogenous oscillators in the brains of the speaker and the listener become mutually entrained on the basis of the speaker's rate of syllable production [13]. Therefore, it seems plausible that neural oscillators might play a role in the production and perception of speech by entraining to the rhythmic properties across various prosodic levels present in the speech signal [1, 6]. Furthermore, these mechanisms can be used to explain synchronization not just of speech, but also of gestures and speech [35]. For example, in one study, subjects were unable to de-synchronize tapping and emphatic stress, showing that these modalities are tightly linked [25].

We do not propose entrainment as the only mechanism for timing backchannels. The dependence of speech timing on semantic and pragmatic factors, as well as the observed temporal irregularities in speech signals, argue against that view. The above mentioned timing by prediction can easily account for temporal irregularities, and is cognitively plausible. However, oscillators are capable of entrainment, and oscillations are present everywhere in the brain, so using them for timing would be a very fast coordinative strategy. Therefore it seems plausible to assume that oscillators play an important role, particularly on smaller time scales. When considering larger time scales, the influences of other cognitive mechanisms, and the irregularities introduced by them, become more apparent. Observations showing that rhythmic patterns are language specific and can be learned [34] also argue for complex interactions between oscillators and other timing mechanisms.

Why is it hard to detect micro-timing?

Detecting micro-timing in the production of backchannels would mean that one could find entrainment of their timing to the timing of certain rhythmic events (e.g., vowel onsets or stressed syllables) in the interlocutor's speech. But as discussed in the previous section, spontaneous speech is not completely regular. Its timing is influenced by physical constraints as well as semantic and pragmatic factors. The relative strengths of these factors might change over time. Besides, perfect synchrony is not expected in natural dialogues. This may make it hard to detect entrainment by statistical methods applied on the whole conversation.

Another difficult question relates to what kind of event in backchannel production is actually entrained to what kind of event in the interlocutor's speech. For example, is a head gesture entrained to pitch accents or vowel onsets? How does this interact with the demonstrated strong correlation to mutual gaze [26, 28]? And does entrainment refer to the onset of a head nod or to its maximum amplitude? Of course, entrainment does not necessarily mean in-phase synchronization. A stable phase offset can also result from entrainment, constituting evidence of temporal coordination.

How can a machine learn micro-timing?

The standard approach to backchannel timing has been to study it empirically in humans, derive some simple rules, and evaluate their impact in studies where humans observe a dialogue between a human speaker and a machine providing backchannels. We have previously conducted a study where we did just that for micro-timing [11]. However, although we found positive effects of the particular macro-timing strategy that we designed to complement micro-timing, we could not find an independent positive effect of micro-timing. We are currently working on improving the study design such that we can study micro-timing in more detail.

Of course, there are overwhelmingly many possible micro-timing strategies, so it is desirable to use machine learning for finding a suitable one. Regarding macro-timing, parasocial consensus sampling (PCS) has recently been proposed as a method to learn opportunities for backchanneling [9]. The basic idea is to show a video of the speaker from a pre-recorded dialogue to subjects who are instructed to provide backchannels just as if they were in real interaction with the speaker. It has been observed that many humans can deal with this parasocial interaction quite well. Time is then discretized in the dialogue, and for each time interval, it is counted how many subjects provided a backchannel. Whenever this number exceeds a predefined threshold, a backchannel is produced. It has been shown that if the resulting behavior is displayed on a conversational agent, it is considered as significantly more natural than a replay of the backchanneling behavior of an individual listener on average. Furthermore, if features like sound amplitude, pitch, or gaze are extracted from the interlocutor's speech immediately before the backchannel opportunity, they can be used to learn a general backchanneling strategy rather successfully [8].

In ongoing studies, we use parasocial consensus sampling for establishing a ground truth about an adequate number

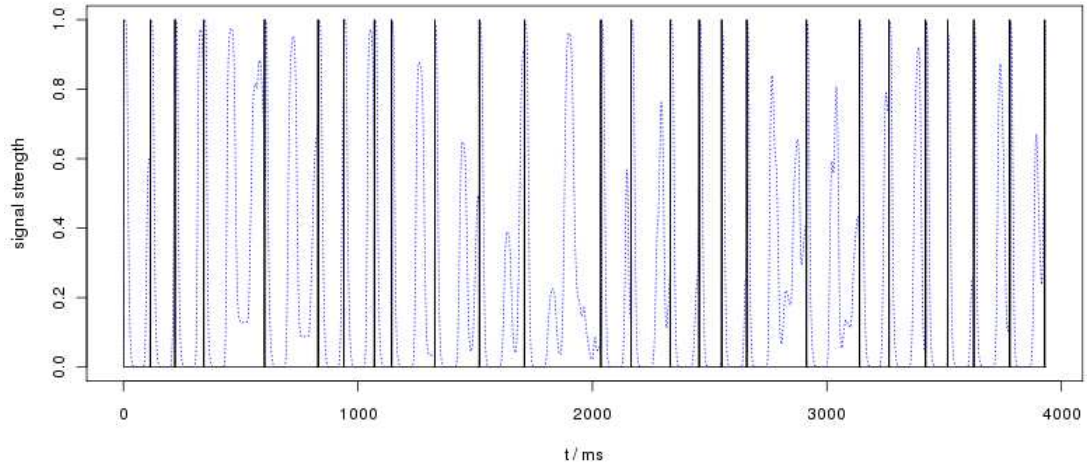

Figure 1: Example output trajectories from an o oscillator bank (dotted, blue) and vowel onsets (black) for syllables in the German phrase "... eine Urlaubsreise mit meiner Familie, also ich war mit meiner Schwester und meiner Mutter dort." ("... a vacation trip with my family, that is, I was there with my sister and my mother."). From [10].

of backchannels during particular segments of conversations. However, we do not use it for learning particular micro-timing rules yet. Instead, we adjust backchannel onsets to rhythmic events in the interlocutor's speech, and/or use an oscillator bank as computational model for the entrained listener that generates candidate timings for backchannels. One reason is that it is not yet clear whether timing in social vs. parasocial interactions is exactly the same. In a recent study, this question has been addressed to some extent using a clever design where a speaker talked simultaneously to three listeners, and perceived only the backchannels from one of them, but the listeners did not know whether they were the ones in real interaction or not [3]. It was found that the numbers of provided backchannels did not differ significantly between the classes of listeners, and that the subjects who were not in real interaction typically did not notice that. However, subjects who watched the videos in a followup experiment were able to identify the real interaction partners with an above chance probability. The study did not address micro-timing issues. That is something we currently study using our PCS corpus.

In the original PCS study, subjects had to press a key when willing to provide a backchannel. This is not suitable for learning micro-timing because it introduces additional delays. In our most recent study [22], we recorded visual and verbal backchannels on video, and subsequently annotated these videos to get more accurate timings. In the future, we plan to use these timings for learning general strategies as well. Instead of discretizing and threshold setting (which can still be used to learn a complementary macro-timing strategy), we will take every backchannel from any subject as a positive instance suitable for learning micro-timing.

Of course, a basic limitation of parasocial consensus sampling is that it cannot capture entrainment of oscillators as a mutual phenomenon. A listener can certainly also influence the speaker's speech rate and other qualities by the frequency and timing of backchannels provided. Such dynamics gets lost in parasocial interactions. Studies are necessary on what additional contribution such mutual entrainment phenomena can make to the perceived naturalness and pleasantness of conversations.

How can micro-timing be implemented and evaluated?

The evaluation of hand-designed or learned strategies for micro-timing poses some difficulties. Firstly, we need to make sure that the used hardware can display micro-timing with sufficient temporal accuracy. Standard LCD monitors with their refresh rate of 60 Hz (i.e., one image every ˜ 17 ms; however, standard videos are encoded at 25 Hz only) and various declared or undeclared delays [4] are not sufficient for distinguishing strategies that differ only by a few milliseconds. However, some old CRT screens, as well as some recent LCD monitors designed for computer gaming, are better suited for these purposes. Of course, if strategies are displayed on a real robot, then gestures will be even slower. They must be planned in advance to be synchronous to speech. Current systems are often not even completely successful at the level of macro-timing in this regard [32]. Of course, given that virtual conversational agents are complex pieces of software, often realized as multiagent architectures, they are often far from being real time as well. One workaround that we used is to create a "flat" version of the virtual agent by recording typical gestures and utterances of the agent, and then generating a video from still images where each frame is reliably timed.

Furthermore, it must be made sure that humans can perceive such slight differences. An action potential of a neuron may already take in the order of 1 ms, but in the case of interaural time differences, humans are sensitive to 10-20 µs [15]. The fusion threshold, beyond which two signals are perceived as one, is typically being reported to be in the range

of 1-3 ms for auditory stimuli, and 20 ms for visual stimuli [20]. Just noticeable differences for visual vs. auditory stimuli are typically around 25-50 ms, with wide variability among different individuals and modes of presentation [14]. The sensitivity to timing differences in dialogues involving feedback is something we still have to investigate.

Finally, micro-timing may be a subtle effect — one that is impossible to perceive if the other aspects of backchannel placement are not right. Therefore, a good macro-timing strategy is a prerequisite for studying micro-timing. In our first study, the effect of differing numbers of backchannels across different strategies was very strong and possibly diluted any effect of micro-timing [11]. Other aspects of behavior like eyeblinks and breathing may also be important to avoid the impression that the conversational agent is staring straight ahead or lifeless. However, these behaviors are rarely recorded or annotated in dialogue corpora. In our study, we used micro-timing for eyeblinks as well (i.e., we synchronized them to vowel onsets) because there is empirical evidence that they are synchronized to speech [19]. As mentioned above, methods like parasocial consensus sampling that are used for learning macro-timing strategies may also not work as well for micro-timing. Obviously, there are many interesting research questions connected to these issues.

Conclusions

Micro-timing is a plausible but hard to grasp feature of timing in dialogues, and possibly other interactions. Studying it not only helps to better understand how humans process rhythms, but also has the potential to make human-robot interaction more natural and pleasant. To do so, new methods for learning and presenting dialog behavior with higher temporal precision have to be developed and used.

Acknowledgments

This research is kindly supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673.

1. REFERENCES

[1] G. Buzsáki and A. Draguhn. Neuronal oscillations in cortical networks. *Science*, 304:1926–1929, 2004.

[2] F. Cummins. Practice and performance in speech produced synchronously. *Journal of Phonetics*, 31:139–148, 2003.

[3] I. de Kok and D. Heylen. Analyzing nonverbal listener responses using parallel recordings of multiple listeners. *Cognitive Processing*, 13:S499–S506, 2012.

[4] T. Elze. Achieving precise display timing in visual neuroscience experiments. *Journal of Neuroscience Methods*, 191:171–179, 2010.

[5] S. Garrod and M. J. Pickering. *Alignment in Communication*, chapter Interactive alignment and prediction in dialogue, pages 193–204. John Benjamins, 2013.

[6] O. Ghitza and S. Greenberg. On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, 66:113–126, 2009.

[7] M. J. Hove and J. L. Risen. It's all in the timing: Interpersonal synchrony increases affiliation. *Social Cognition*, 27:949–960, 2009.

[8] L. Huang, L.-P. Morency, and J. Gratch. Learning backchannel prediction model from parasocial consensus sampling: A subjective evaluation. In *Proceedings of the International Conference on Intelligent Virtual Agents*, 2010.

[9] L. Huang, L.-P. Morency, and J. Gratch. Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. In *Proceeedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, 2010.

[10] B. Inden, Z. Malisz, P. Wagner, and I. Wachsmuth. Rapid entrainment to spontaneous speech: A comparison of oscillator models. In *Proceedings of the Cognitive Science Conference*, 2012.

[11] B. Inden, Z. Malisz, P. Wagner, and I. Wachsmuth. Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent. In J. Epps, F. Chen, S. Oviatt, K. Mase, A. Sears, and K. Jokinen, editors, *Proceedings of the 15th International Conference on Multimodal Interaction, ICMI'13 - Sydney*, 2013.

[12] I. Jauk, P. Wagner, and I. Wachsmuth. Dynamic perception-production oscillation model in human-machine communication. In *Proceedings of the International Conference on Multimodal Interaction*, 2011.

[13] M. Kawasaki, Y. Yamada, Y. Ushiku, E. Miyauchi, and Y. Yamaguchi. Inter-brain synchronization during coordination of speech rhythm in human-to-human social interaction. *Nature*, 2013.

[14] M. Keetels and J. Vrommen. *The neural bases of multisensory processes*, chapter Perception of synchrony between the senses, pages 147–177. CRC Press, 2011.

[15] R. G. Klumpp and H. R. Eady. Some measurements of interaural time difference thresholds. *Journal of the Acoustic Society of America*, 28:859, 1956.

[16] S. Kopp, J. Allwood, K. Grammer, E. Ahlsen, and T. Stockmeier. Modeling embodied feedback with virtual humans. In I. Wachsmuth and G. Knoblich, editors, *Modeling communication with robots and virtual humans*. Springer-Verlag Berlin Heidelberg, 2008.

[17] E. W. Large. *Dynamic Representation of Musical Structure*. PhD thesis, The Ohio State University, 1994.

[18] E. W. Large, F. V. Almonte, and M. J. Velasco. A canonical model for gradient frequency neural networks. *Physica D*, 239:905–911, 2010.

[19] D. Loehr. Aspects of rhythm in gesture and speech. *Gesture*, 7:179–214, 2007.

[20] M. Lotze, M. Wittmann, N. von Steinbüchel, E. Pöppel, and T. Roenneberg. Daily rhythm of temporal resolution in the auditory system. *Cortex*, 35:89–100, 1999.

[21] R. M. Maatman, J. Gratch, and S. Marsella. Natural behavior of a listening agent. In *Proceedings of the International Conference on Interactive Virtual Agents*, pages 25–36, 2005.

[22] Z. Malisz, B. Inden, P. Wagner, and I. Wachsmuth. Timing and entrainment of multimodal feedback in dialogue. A simulation with an embodied conversational agent. *submitted*, 2014.

[23] Z. Néda, E. Ravasz, T. Vicsek, Y. Brechet, and A. L. Barabasi. The physics of rhythmic applause. *Physical Review E*, 61:6987, 2000.

[24] M. O'Dell and T. Nieminen. Coupled oscillator model of speech rhythm. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, 1999.

[25] B. Parrell, L. Goldstein, S. Lee, and D. Byrd. Temporal coupling between speech and manual motor actions. In *Proceedings of the Ninth International Seminar on Speech Production.*, 2011.

[26] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi. A model of attention and interest using gaze behavior. In *Proceedings of the 5th International Working Conference on Intelligent Virtual Agents.*, 2005.

[27] M. J. Pickering and S. Garrod. Towards a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226, 2004.

[28] R. Poppe, K. P. Truong, D. Reidsma, and D. Heylen. Backchannel strategies for artificial listeners. In *Proceedings of the Intelligent Virtual Agents Conference*, 2010.

[29] C. Richardson, R. Dale, and K. Schockley. Synchrony and swing in conversation: Coordination, temporal dynamics, and communication. In I. Wachsmuth, M. Lenzen, and G. Knoblich, editors, *Embodied Communication in Humans and Machines*. Oxford University Press, 2007.

[30] G. Rizzolatti and M. A. Arbib. Language within our grasp. *Trends in Neurosciences*, 27:169–192, 1998.

[31] M. Salem, S. Kopp, I. Wachsmuth, and F. Joublin. A multimodal scheduler for synchronized humanoid robot gesture and speech. In E. Efthimiou and G. Kouroupetroglou, editors, *9th International Gesture Workshop (GW2011)*, pages 64–67, 2011.

[32] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4:201–217, 2012.

[33] I. Wachsmuth and S. Kopp. Lifelike gesture synthesis and timing for conversational agents. In I. Wachsmuth and T. Sowa, editors, *International Gesture Workshop, GW 2001, London*, volume 2298 of *LNAI*, pages 120–133. Springer-Verlag, 2002.

[34] P. Wagner, Z. Malisz, B. Inden, and I. Wachsmuth. Interaction phonology — a temporal co-ordination component enabling communicative alignment. In *Towards a New Theory of Communication*. John Benjamins, in press.

[35] P. Wagner, Z. Malisz, and S. Kopp. Gesture and speech in interaction: an overview. *Speech Communication*, 57:209–232, 2014.

[36] M. Wilson and T. P. Wilson. An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review*, 12:957–968, 2005.

[37] M. Włodarczak, J. Simko, and P. Wagner. Temporal entrainment in overlapped speech: Cross-linguistic study. In *Proceedings of Interspeech*, 2012.

[38] M. Włodarczak, J. Simko, and P. Wagner. Pitch and duration as a basis for entrainment of overlapped speech onsets. In *Proceedings of Interspeech*, 2013.

[39] M. Włodarczak, J. Simko, P. Wagner, M. O'Dell, M. Lennes, and T. Nieminen. Finnish rhythmic structure and entrainment in overlapped speech. In *Nordic Prosody. Proceedings of the XIth Conference*, 2013.

[40] D. M. Wolpert and M. Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, 11:1317–1329, 1998.