# Research focus: Interactional aspects of spoken face-to-face communication

*Jonas Beskow, Jens Edlund, Joakim Gustafson, Mattias Heldner, Anna Hjalmarsson, David House*
*KTH Speech, Music and Hearing*

***Abstract***

*We have a visionary goal: to learn enough about human face-to-face interaction that we are able to create an artificial conversational partner that is human-like. We take the opportunity here to present four new projects inaugurated in 2010, each adding pieces of the puzzle through a shared research focus: interactional aspects of spoken face-to-face communication.*

## Introduction

We have formulated a visionary goal: *to learn enough about human face-to-face interaction that we are able to create an artificial conversational partner that is human-like,* or as Cassell (2007) puts it, ”acts human enough that we respond to it as we respond to another human”.

Note that this is not a case of implementing science fiction – it is about testing scientific hypotheses. Paraphrasing the methods used in cognitive science, we implement models of human behaviour and put these to the test in interactions with humans. This is how we validate our understandings.

While our dream of human-likeness may eventually turn out to be overly ambitious, it has already been useful to us by guiding our research efforts towards black holes in our knowledge about human conversation. It is for example evident that state-of-the-art speech technology neither sounds like a conversational partner, nor understands fundamental aspects of human conversational behaviour.

A massive amount of cross-disciplinary research is needed to realize the visionary goal. Although we have a long way to go, we learn useful lessons from every step. We are currently involved in a half-dozen national and international research projects orchestrated towards our goal.

In this paper, we take the opportunity to give a brief overview of four new projects with national funding sharing a common research focus: *interactional aspects of spoken face-to-face communication*.

## Background

Speech in conversation differs from speech in situations where there is no conversational partner present, for example read-aloud speech. This becomes apparent when listening to people acting or reading dialogues aloud. We can often tell that a conversation is acted, even if the script is meticulously written with all imaginable detail. Something special happens when there is someone to interact with, and what is being said is planned there and then.

Ironically, we know less about the primary use of speech – in face-to-face communication – than about many other kinds of speech (see e.g. Heldner & Edlund, 2007). This is, among other things, reflected in how the talking computers we encounter today behave. For one thing, they simply do not sound like they are having a conversation even if they say the same thing a human in the same situation would. In addition, face-to-face conversation involves other sources of information, perhaps most notably a visual channel with gaze, nods, other gestures, posture, proxemics etc. that forms an intrinsic part of the communication. Thus, further investigations about how humans converse are motivated from basic and applied research perspectives alike.

An initial requirement for making substantial progress is access to spontaneous conversations. We have recently collected about 60 hours of audio, video and motion capture data in conversations within the project *Spontal: Multimodal database of spontaneous speech in dialog* (Edlund, et al., 2010). The projects presented here all explore this dataset.

# Current projects

The following is a brief overview of four new projects within the research theme interactional aspects of spoken face-to-face communication.

## Prosody in conversation

Riksbankens Jubileumsfond (RJ) has granted the project *Prosody in conversation (Samtalets prosodi)* 5.2 MSEK for the years 2010-2012 (contract P09-0064:1-E). Applicant: Mattias Heldner.

The project investigates how people talking to each other jointly decide *who should speak when*, and the role of prosody in making these joint decisions. While prosody is by no means the only relevant information for this joint interaction control, we believe that it plays a central role (see e.g. Edlund & Heldner, 2005, and references mentioned therein). A detailed model of the prosody involved in interaction control is crucial both for *producing* appropriate conversational behaviour and for *understanding* human conversational behaviour. Both are required in order to reach our visionary goal, and represent the artificial conversational partner in the roles of speaker and listener in a conversation, respectively.

One line of inquiry within the project is the quantitative acoustic analysis of prosodic features in genuine spoken face-to-face conversations. The project focuses on local intonation patterns in the immediate vicinity of interactional events, such as transitions from (i) speech to pauses (within-speaker silences); (ii) speech to gaps (between-speaker silences, i.e. at speaker changes); and (iii) speech by one speaker to speech by another speaker (i.e. overlapping speech in speaker changes). In addition, we analyze selected interactional phenomena occurring on a longer time scale, such as pitch similarity across these interactional events and the overall tendency of interlocutors to grow increasingly similar as the conversation proceeds. This increasing interlocutor similarity reported in the literature under many names (e.g. entrainment, alignment, accommodation; see e.g. Edlund, Heldner, & Hirschberg, 2009 for an overview) has been reported for a great number of linguistic features, but we limit ourselves to prosody in this project.

In addition, the results of the acoustic analyses are fed into a second line of inquiry: studies of the effects of using or introducing such prosodic features in a conversation. These studies will include listening experiments where manipulations of genuine conversations by means of re-synthesis are used as stimuli. Furthermore, there will be pragmatic experiments where the conversational behaviour in response to the use of such prosodic features in artificial speech is analyzed. Finally, there will be analyses of conversational behaviour in response to real-time (or minimum delay) manipulations of genuine conversations, such as deletions, insertions or manipulations of features.

## The rhythm of conversation

The Swedish Research Council (VR) HS, has granted the project *Rhythm of conversation (Samtalets rytm)* 2.9 MSEK for the years 2010-2012 (contract 2009-1766). Applicant: Mattias Heldner.

The project *Rhythm of conversation* investigates how a set of rhythmic prosodic features contributes to the joint interaction control in conversations. Of particular interest is acoustic descriptions of features related to variations in speech rate (i.e. accelerations and decelerations in speech rate) and loudness (i.e. increases and decreases in loudness), and how these are used for interactional purposes.

Loudness is generally perceived as an important component in the signalling of prosodic functions such as prominence and boundaries (cf. Lehiste & Peterson, 1959). Attempts to capture this impression in acoustic analyses, however, regularly show only weak correlations with these functions (e.g. Fry, 1955; Lieberman, 1960). This has lead much prosodic research to concentrate on melodic prosodic aspects – sometimes to the extent that prosody is equated with pitch. Recent work indicates, however, that loudness may be a strong correlate of such functions, when measured as subjective loudness rather than as physical intensity (Kochanski, Grabe, Coleman, & Rosner, 2005). This is highly unexplored and something we pursue in connection with rhythm as an interactional phenomenon.

We want to find out, for example, whether the speech rate and loudness variations (prosodic features that are complementary to those studied in *Prosody in conversation*) before pauses (i.e. within-speaker silences) are different from those before gaps (between-speaker silences), or whether they display differences before backchannel-like utterances compared to other utterances.

## Introducing interactional phenomena in speech synthesis

The Swedish Research Council (VR) NT, has granted the project *Introducing interactional phenomena in speech synthesis (Talsyntes för samtal)* 2.1 MSEK for the years 2010-2012 (contract 2009-4291). Applicant: Joakim Gustafson.

The project recreates human interactional vocal behaviour in speech synthesis in three phases. The first deals with what Allwood (1995) calls feedback morphemes and Ward (2000) conversational grunts (e.g. mm, eh). We also include audible breathing, following Local & Kelly (1986) who hold breath as a strong interactional cue. These tokens are traditionally missing in speech synthesis. We remedy this by (1) annotating instances of them in the Spontal corpus (and possibly other corpora), (2) synthesizing the missing tokens using several methods, and (3) evaluating the results in a series of experiments comparing synthesized versions with the originals as well as evaluating their perceived meaning and function.

The second phase is similarly structured, but targets events that occur in the transitions between speech and silence and back – transitions that vary depending on the situation. We focus on three transition types: *normal* (the target of current syntheses), *hesitant* and *abrupt*. Pauses and retardations are strong cues for hesitation, and glottal stops or unreleased plosives are frequently a result of an interruption (Local & Kelly, 1986). Speech that has been halted on a glottal stop or an unreleased plosive can be restarted by releasing the stop. This gives rise to different acoustic effects which we recreate and evaluate.

In the third phase, we evaluate reactions to a dialogue system making use of the synthesized cues developed in the first two phases. In semi-automatic dialogue systems modelling speaking and listening as parallel and mutually aware processes, we use two scenarios to verify and validate our results: the attentive speaker – an interruptible virtual narrator making use of synthesized cues for hesitation and end-of-contribution; and the active listener – an information gathering system, aiming to encourage the user to continue speaking (cf. Gustafson, Heldner, & Edlund, 2008).

## Intonational variation in questions in Swedish

The Swedish Research Council (VR) HS, has granted the project *Intonational variation in questions in Swedish (Variation i frågeintonation i svenska)* 2.6 MSEK for the years 2010-2012 (contract 2009-1764). Applicant: David House.

The project investigates and describes phonetic variation of intonation in questions in spontaneous Swedish conversation, with an initial premise that there does not exist a one-to-one relationship between intonation and sentence type (Bolinger, 1989). The Spontal database is used to find a general understanding of the role of questions in dialogue and an explanation of why descriptions of question intonation has proven so difficult. We expect to find certain patterns of intonation that correlate with for example dialogue and social function.

We will test several hypotheses from the literature. One example is the hypothesis that there is a larger proportion of final rises and high pitch in questions which are social in nature than in those which are information oriented. Another example concerns the type of visual gestures that accompany questions (McNeill, 1992): we will test the hypothesis that iconic and emblematic gesture types co-occur more often with information-oriented questions while dialogue gestures such as nods and facial gestures co-occur more frequently with social-oriented questions.

Finally, our results will be analyzed within the framework of biological codes for universal meanings of intonation proposed by Gussenhoven (2002). Gussenhoven describes three codes, or biological metaphors: a frequency code, originally proposed by Ohala (1983), implying that a raised F0 is a marker of submissiveness or non-assertiveness and hence question intonation; an effort code, in which higher F0 requires increased articulation effort which highlight important focal information; and a production code associating high pitch with phrase beginnings (new topics) and low pitch with phrase endings. A biological explanation for the pragmatic functions of intonation is of important theoretical interest for the project, and leads further into investigations of the relationships between intonation and visual gestures in a framework of biological codes.

# Summary

We have proposed an ambitious and visionary goal for our research: to learn enough about human face-to-face interaction that we are able to create an artificial conversational partner that is human-like in the sense that people interacting with it respond to it as they do to other humans. This visionary goal has been instrumental in the prioritization and formulation of a current research focus for our group: investigations of interactional aspects of spoken face-to-face communication. We have described four new externally funded projects that are representative of and will advance the research frontier within this common research focus.

While these projects do not in themselves have either the resources or the scope to reach our visionary goal, they each add a piece of the puzzle, and we are confident that they will help identify future areas for research contributing towards the long-term goal. The visionary goal requires a wider scoped platform for future grant applications. The joint effort of these projects forms a strong point of departure by providing critical mass of research expertise in the area.

# Acknowledgements

This research is carried out at KTH Speech, Music and Hearing. Funding was provided by Riksbankens Jubileumsfond (RJ) project P09-0064:1-E *Prosody in conversation;* the Swedish Research Council (VR) projects 2009-1766 *The rhythm of conversation*, 2009-4291 *Introducing interactional phenomena in speech synthesis*; 2009-1764 *Intonational variation in questions in Swedish*; and 2006-7482 *Spontal: Multimodal database of spontaneous speech in dialog*.

# References

Allwood, J (1995). An activity based approach to pragmatics. In: H Bunt & B Black, eds, *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*. Amsterdam, The Netherlands: John Benjamins.

Bolinger, D (1989). *Intonation and its uses: Melody in grammar and discourse*. London, UK: Edward Arnold.

Cassell, J (2007). Body language: Lessons from the near-human. In: J Riskin, ed, *Genesis Redux: Essays in the History and Philosophy of Artificial Life*. Chicago, IL, USA: The University of Chicago Press, 346-374.

Edlund, J, Beskow, J, Elenius, K, Hellmer, K, Strömbergsson, S, & House, D (2010). Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In *Proceedings of LREC 2010*. Valetta, Malta.

Edlund, J, & Heldner, M (2005). Exploring prosody in interaction control. *Phonetica,* 62: 215-226.

Edlund, J, Heldner, M, & Hirschberg, J (2009). Pause and gap length in face-to-face interaction. In *Proceedings of Interspeech 2009*. Brighton, UK, 2779-2782.

Fry, D B (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America,* 27: 765-768.

Gussenhoven, C (2002). Intonation and interpretation: phonetics and phonology. In: B Bel & I Marlien, eds, *Proceedings of the Speech Prosody 2002 Conference*. Aix-en-Provence, France, 47-57.

Gustafson, J, Heldner, M, & Edlund, J (2008). Potential benefits of human-like dialogue behaviour in the call routing domain. In *Perception in Multimodal Dialogue Systems*. Berlin, Germany: Springer, 240-251.

Heldner, M, & Edlund, J (2007). What turns speech into conversation? A project description. In *TMH-QPSR 50: Fonetik 2007*. Stockholm, Sweden, 45-48.

Kochanski, G, Grabe, E, Coleman, J, & Rosner, B (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America,* 118: 1038-1054.

Lehiste, I, & Peterson, G E (1959). Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America,* 31: 428-435.

Lieberman, P (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America,* 32: 451-454.

Local, J K, & Kelly, J (1986). Projection and 'silences': Notes on phonetic and conversational structure. *Human Studies,* 9: 185-204.

McNeill, D (1992). *Hand and mind – What gestures reveal about thought*. Chicago, IL, USA: University of Chicago Press.

Ohala, J J (1983). Cross-language use of pitch: an ethological view. *Phonetica,* 40: 1-18.

Ward, N (2000). The challenge of non-lexical speech sounds. In *Proceedings of ICSLP 2000*. Beijing, China, 571-574.