What turns speech into conversation? A project description

Mattias Heldner & Jens Edlund
KTH Speech, Music and Hearing

Abstract

The project Vad gör tal till samtal? (What turns speech into conversation?) takes as its starting point that while conversation must be considered the primary kind of speech, we are still far better at modelling monologue than dialogue, in theory as well as for speech technology applications. There are also good reasons to assume that conversation contains a number of features that are not found in other kinds of speech, including, among other things, the active cooperation among interlocutors to control the interaction, and to establish common ground. Through this project, we hope to improve the situation by investigating features that are specific to human-human conversation – features that turns speech into conversation. We will focus on acoustic and prosodic aspects of such features.

Introduction

A long line of research in phonetics and speech technology has given us basic knowledge of how speech works in communication. We have a good picture of what the building blocks of speech are; how parts of the speech signal are made more prominent; how stretches of speech are grouped and delimited; and of the form and function of these components.

However, the object of study in this research has predominantly been taken from situations where there is not any interaction between speakers and listeners. It has typically dealt with isolated words, isolated utterances, read-aloud speech, monologue, computer directed speech, and so on; and more seldom with conversations among humans. This is so despite the fact that face-to-face conversation must be considered the primary and the richest kind of speech. Speech and language originates in conversations; this is the situation in which we learn to speak; and conversation is the most natural way of communicating for most of us. There are also good reasons to assume that *speech in conversation* differs significantly from speech in situations where no listener is present. Among other things there is a control of the interaction in conversations that is lacking in other kinds of speech – a characteristic of conversations we have elsewhere referred to as *interaction control* (e.g. Edlund & Heldner, 2005; Heldner, Edlund, & Carlson, 2006).

We argue that the choice of methods and materials in our field has resulted in a situation where the knowledge about speech in conversation is to a large extent lacking, and that we are much better at modelling monologue than dialogue – in theory as well as for practical applications. The goal of this project is to improve the situation by investigating and modelling some of the phenomena that turn speech into conversation – the continuous collaboration around turn-taking and on establishing *common ground*. Since there is substantial evidence that prosody is an important component in interaction control, we will concentrate on modelling prosody for interaction control, and on investigating human reactions to behaviour generated from those models.

The project is primarily intended to shed light on the way humans communicate with speech. We will do this by building a model of human conversation piece-by-piece. Apart from advancing fundamental research, such a model may also be used for more practical applications, such as for improving tools for communication between humans, or for building spoken interfaces to computers that better match the expectations on what spoken conversation is supposed to be like.

In this contribution, we will present our new project, our research questions, and outline three areas within which we will perform investigations.

Background

A basic condition for conversation is at least two parties that are able as well as willing to talk to each other. During the conversation they take on the roles as speaker and listener, and the roles recurrently change during the course of the conversation. This changing of roles, usually referred to as *turn-taking* (e.g. Cassell, Bickmore, Campbell, Vilhjámsson, & Yan, 2000; Goodwin, 1981; Sacks, Schegloff, & Jefferson, 1974), is unique to conversation. Several research questions in this project are formulated from the perspectives of the speaker and the listener in a conversation. The topics we will look into include:

- What the listener does to find suitable places to speak, or to indicate that (s)he wants to speak.
- What the speaker does to (indicate that (s)he wants to) keep the floor, or to hand it over to someone else.

Furthermore, the project takes as a starting point that conversation is a collaboration between the interlocutors, and that both speaker and listener actively and continuously contributes to the conversation. The speaker, obviously, by saying something. The listener by giving feedback to the speaker on different levels, including that listener and speaker have established contact; that the listener has perceived and understood what was said; whether the listener has accepted what was said; as well as other attitudes towards what was said (e.g. Allwood, Nivre, & Ahlsén, 1993; Clark, 1996). This feedback behaviour is also unique to conversation. In relation to this, we want to look into the following topics:

- What the listener does to find suitable places to give verbal feedback.
- How verbal feedback on different levels is signalled?

But there is so much more going on in a conversation. For example, before the speaker can initiate the conversation (s)he must get the listeners' attention; the speaker as well as the listener may indicate that they want to end the conversation etc. These are also potential objects of study in the project. To complicate matters further, speech is naturally not the only ingredient in a conversation. All of the above mentioned functions could likely be performed with other means, such as nods, gaze, or gestures, as well. In this project, however, we will limit ourselves to investigations of the acoustic and/or prosodic channels that are relevant to conversation.

Methods

A brief note on methods in the project. As mentioned above, we will concentrate on acoustic/prosodic features of phenomena that are specific to conversations. Furthermore, we will study such features with operationally defined concepts, analyses and judgments. To the extent possible, we will use automatic instrumental methods and observations to ensure that the results can be replicated. Whenever humans are used as judges, we will use statistical tests to ensure that the agreement and reliability figures of their judgments are sufficiently high.

For example, instead of choosing measurement points manually we will use explicitly defined criteria to extract acoustic/ prosodic features automatically. Similarly, the analysis of conversational behaviour will to a large extent rely on an analysis of conversational states and transitions between those states that can also be extracted automatically (see e.g. Jaffe & Feldstein, 1970). Such methods will also enable us to examine larger samples.

Automatically extracted prosodic features as a means of studying or modelling conversational phenomena have successfully been used by a number of research groups in the past (e.g. Cassell et al., 2000; Ferrer, Shriberg, & Stolcke, 2002; Shriberg & Stolcke, 2004; Shriberg, Stolcke, Hakkani-Tür, & Tür, 2000; Ward, 1999; Ward & Tsukahara, 2000).

Three areas of research

We are planning to explore three areas of research within the project. First, we want to approach the question of where to look for interaction control signals from a new angle. The core of the project will then be comprised of investigations of acoustic or prosodic features occurring in connection with interaction control phenomena. In the final phase, to test whether the features we have included in our models have any perceptual relevance. These areas will be described more in detail below.

Where to look for interaction control signals?

There is research suggesting that any turn-yielding signals (i.e. an indication that it is suitable for someone other than the current speaker to say something) have to occur at least 200-300ms before the next speaker starts because of the minimal response times for spoken utterances (e.g. Wesseling & van Son, 2005a, 2005b). Turn-keeping signals (i.e. an indication that the speaker is not yet finished despite a disruption of the flow of speech) may conceivably occur later than the turn-yielding signals, as they are not meant to trigger a response, but rather to inhibit one.

Such observations renders investigations of actual silence durations between utterances in conversations highly interesting, as they ought to give an indication of where to look for potential interaction control signals.

It is often claimed that human turn-taking is so precise that next speakers tend to start with *no gap* and *no overlap.* This claim is in turn often used to support the additional claims that turn-taking must rely solely on the ability to *project* (in the sense of anticipating) upcoming turn ends, and that this projection is based solely on syntactic (and absolutely not on prosodic) information (e.g. de Ruiter, Mitterer, & Enfield, 2006; Levinson, 1983). However, neither our own preliminary observations, nor published quantitative data seem to support the view of zero gap between turns. Rather it seems that the median of such silence distributions comes closer to some 300ms (e.g. ten Bosch, Oostdijk, & Boves, 2005; Weilhammer & Rabold, 2003), thus opening the possibility that prosodic information before the silence might still be relevant for turn-taking. We do not consider projection and prosodic turn-ending signals as mutually exclusive, but would rather stress that redundancy is a well-studied and recurring feature of language. We have also seen some preliminary evidence that projection and prosodic turn-ending are used for slightly different purposes. Firstly and foremost, however, we want to corroborate the idea of using prosodic information for interaction control, and distribution analyses of inter- and intra-speaker silences for a substantial amount of conversational data are underway.

What acoustic or prosodic features occur in connection with interaction control phenomena?

Own research and that of others have shown that the presence of a silence is not sufficient to determine if a speaker has finished what (s)he was going to say or not (cf. e.g. Edlund & Heldner, 2005; Ferrer et al., 2002). In fact, it seems that silences are just as frequent in situations where the speaker has not finished, for example in connection with hesitations or semantically heavy words. Such observations led us to investigate whether other prosodic features may be utlized for making interaction control decisions. So far, we have looked at intonation patterns immediately before silences. This study showed that level tones in the middle of the speaker's F0 range often occured in situations where the speaker was not finished, and that low and falling patterns tended to occur the speaker was actually finished. High and rising patterns, however, were as frequent when the speaker had finished as when (s)he had not (Edlund & Heldner, 2005). While these results seem promising, there is clearly a need for more research on intonation patterns in connection with interaction control.

A number of other prosodic features have also been suggested as relevant for interaction control. One of them is lengthening patterns before prosodic boundaries. As a part of our investigation of intonation patterns, we have developed a method for automatic segmentation into "pseudo-syllables" or "psyllables", mainly comprising the voiced part of the syllable nucleus. We intend to investigate whether the duration of such psyllables before silences may be used to estimate lengthening patterns, and in turn whether such lengthening patterns may be useful for making interaction control decisions.

In addition, we intend to explore a suggestion by Local & Kelly (1986) that different vocal tract configurations may be associated with the speaker being finished (open vocal tract) or not finished (closed vocal tract). Such vocal tract configurations could possibly result in different acoustic qualities of silent pauses so that silences where the speaker is finished and exhales should have a higher intensity than those where (s)he is not. It remains to be shown, however, that such differences are detectable and whether they are of any practical use for determining whether the speaker is finished or not.

Another investigation we plan to undertake concerns whether it is possible to generalise the finding that turn keeping is signalled with a level intonation pattern (i.e. no change in intonation) to other kinds of "no change" on an acoustic level. A closed, or frozen vocal tract during silences would be one such "no change" feature, other potential features include prolonged speech sounds resulting in "no changes" in the spectral domain, phonological processes spanning silences and so on.

Are these features perceptually relevant?

The investigations of where to look for interaction control signals and what to look for form the basis of models of human interaction control behaviour. In the final phase of the project, we will investigate whether the features we have included in our models have the expected effect. We will generate interaction control behaviour from the models and observe subjects reactions and behaviours towards such behaviour. For example, will subjects avoid taking the floor when a computer displays turn-keeping signals, and, actually grab the floor in the presence of turn-yielding signals? It is important to note here speech technology provides an opportunity to test models and theories empirically, but that technology is not a goal in itself in this project.

Conclusions

It is our hope and intention that this project will yield significant contributions to the knowledge about what turns speech into conversation by describing and modelling prosodic behaviour relevant to interaction control.

Acknowledgment

The project *Vad gör tal till samtal?* is funded by the Swedish Research Council VR, area for research in Humanities and Social Sciences, from 2007 until 2009. The research that led up to the project was funded by the European Commission's Sixth Framework Program project IP-506909 (CHIL).

References

Allwood J, Nivre J & Ahlsén E (1993). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics, 9:* 1-26.

Cassell J, Bickmore T, Campbell L, Vilhjámsson H & Yan H (2000). Human conversation as a system framework: Designing embodied conversational agents. In: Cassell J, Sullivan J, Prevost S & Churchill E eds, *Embodied Conversational Agents.* Cambridge, M.A.: The MIT Press, 29-63.

Clark H H (1996). *Using language*. Cambridge: Cambridge University Press.

de Ruiter J P, Mitterer H & Enfield N J (2006). Predicting the end of a speaker's turn: a cognitive cornerstone of conversation. *Language, 82:* 515-535.

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

Ferrer L, Shriberg E & Stolcke A (2002). Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody in human-computer dialog. In: *Proceedings of ICSLP 2002.* Denver, USA, 2061-2064.

Goodwin C (1981). *Conversational organization: Interaction between speakers and hearers*. New York: Academic Press.

Heldner M, Edlund J & Carlson R (2006). Interruption impossible. In: *Nordic Prosody: Proceedings of the IXth Conference, Lund 2004.* Frankfurt am Main: Peter Lang, 97-105.

Jaffe J & Feldstein S (1970). A descriptive classification of conversational rhythms. In: *Rhythms of Dialogue.* New York: Academic Press, 9-27.

Levinson S C (1983). *Pragmatics*. Cambridge: Cambridge University Press.

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

Sacks H, Schegloff E A & Jefferson G (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50:* 696-735.

Shriberg E & Stolcke A (2004). Direct Modeling of Prosody: An Overview of Applications in Automatic Speech Processing. In: *Proceedings of Speech Prosody 2004* Nara, Japan, 575-582.

Shriberg E, Stolcke A, Hakkani-Tür D & Tür G (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication, 32:* 127-154.

ten Bosch L, Oostdijk N & Boves L (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication, 47:* 80-86.

Ward N (1999). Low-pitch regions as dialog signals? Evidence from dialog-act and lexical correlates in natural conversation. In: *ESCA Workshop on Dialog and Prosody.* Eindhoven: TUE-IPO, 83-88.

Ward N & Tsukahara W (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics, 32:* 1177-1207.

Weilhammer K & Rabold S (2003). Durational aspects in turn taking. In: *Proceedings of ICPhS'03.* Barcelona, Spain,

Wesseling W & van Son R J J H (2005a). Early preparation of experimentally elicited minimal responses. In: *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue.* Lisbon, Portugal, 11-18.

Wesseling W & van Son R J J H (2005b). Timing of experimentally elicited minimal responses as quantitative evidence form the use of intonation in projecting TRPs. In: *Proceedings of Interspeech'2005.* Lisbon, Portugal, 3389-3392.