points that – as far as possible – provide trustworthy formant values for all recordings. The other is a script which by tracing the intensity variation in each partial as the f0 changes, can be used to determine when a given partial crosses a formant. By measuring f0 at this point and counting the number of the partial we are able to estimate the formant frequency. We assume that this approach will be more accurate than judging the formant frequencies by visual inspection alone.

It is obvious that the "f0-sweep" approach we use to determine formant values manually is not without flaws as we are relying heavily on a number of assumptions: First we expect our speakers to be able to produce the same vowel quality independent of pitch. As vowel quality and pitch are known to be interrelated in real speech it may both be difficult for our speakers to live up to this expectation, and difficult for us to verify auditorily whether they do. Even if the speakers may succeed in 'freezing' the oral cavities during the sweep, differences may arise due to movement of the larynx as the pitch is changed, as well as due to changes in voice quality associated with the pitch. Notably the voice often seemed to get more breathy and hypofunctional towards the lower end of the pitch range. The method of determining the time of the maximum energy for the partial may also be affected by overall changes in intensity that have nothing to do with the interaction between the partial and the formant. This would mostly affect estimates of F1 as the transition of partials through higher formants happens much faster, and since there are often more partials crossing through the formant thus giving more estimates. Finally the accuracy of course depends on the accuracy of the f0 tracing, and more so the higher the partial. Despite the potential shortcomings of the method it does seem to provide reliable results, and is particularly helpful in determining the formant frequencies in the lower region of the spectrum.

Our ongoing analyses of the data have so far only confirmed the usefulness of carrying out the larger investigation. We hope to be able to ensure that our colleagues at the LANCHART Project need not end up reporting as sound changes what might merely be the results of microphone changes...

References

Brixen, E.B., 1996. Spectral Degradation of Speech Captured by Miniature Microphones Caused by the Microphone Placing on Persons' Head and Chest. *Proceedings AES 100th Convention.*

Brixen, E.B., 1998. Near Field Registration of the Human Voice: Spectral Changes due to Positions. *Proceedings AES 104th Convention.*

Plichta, B., 2004. Data acquisition problems. In B. Plichta, *Signal acquisition and acoustic analysis of speech.* Available at: http://bartus.org/akustyk/signal_aquisition.pdf.

van Son, R.J.J., 2002. *Can standard analysis tools be used on decompressed speech?* Available at: http://www.fon.hum.uva.nl/Service/IFAcorpus/SLcorpus/AdditionalDocuments/CoCOSDA2002.pdf.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 53–56

Prosodic Cues for Interaction Control in Spoken Dialogue Systems

Mattias Heldner and Jens Edlund

Department of Speech, Music and Hearing, KTH, Stockholm
{mattias|edlund}@speech.kth.se

Abstract

This paper discusses the feasibility of using prosodic features for interaction control in spoken dialogue systems, and points to experimental evidence that automatically extracted prosodic features can be used to improve the efficiency of identifying relevant places at which a machine can legitimately begin to talk to a human interlocutor, as well as to shorten system response times.

1 Introduction

All spoken dialogue systems, no matter what flavour they come in, need some kind of interaction control capabilities in order to identify places where it is legitimate to begin to talk to a human interlocutor, as well as to avoid interrupting the user. Most current systems rely *exclusively* on silence duration thresholds for making such interaction control decisions, with thresholds typically ranging from 500 to 2000 ms (Ferrer, Shriberg & Stolcke, 2002; Shriberg & Stolcke, 2004). Such an approach has several drawbacks, both from the point of view of the user and that of the system. Users generally have to wait longer for responses than in human-human interactions; at the same time they run the risk of being interrupted by the system, since people frequently pause *mid-speech*, for example when hesitating or before semantically heavy words (Edlund & Heldner, 2005; Shriberg & Stolcke, 2004); and using silent pauses as the sole information for segmentation of user input is likely to impair the system's speech understanding, as unfinished or badly segmented utterances often are more difficult to interpret (Bell, Boye & Gustafson, 2001).

Humans are very good at discriminating the places where their conversational partners have finished talking from those where they have not – accidental interruptions are rare in conversations. Apparently, we use a variety of information to do so, including numerous prosodic and gestural features, as well as higher levels of understanding, for example related to (in)completeness on a structural level (e.g. Duncan, 1972; Ford & Thompson, 1996; Local, Kelly & Wells, 1986).

In light of this, the interaction control capabilities of spoken dialogue systems would likely benefit from access to more of this variety of information – more than just the duration of silent pauses. Ultimately, spoken dialogue systems should of course be able to combine all relevant and available sources of information for making interaction control decisions. Attempts have been made at using semantic information (Bell, Boye & Gustafson, 2001; Skantze & Edlund, 2004), prosodic information and in particular intonation patterns (Edlund & Heldner, 2005; Ferrer, Shriberg & Stolcke, 2002; Thórisson, 2002), and visual information (Thórisson, 2002) to deal with (among other things) the problems that occur as a result of interaction control decisions based on silence only.

2 Prosodic cues for interaction control

Previous work suggests that a number of prosodic or phonetic cues are liable to be relevant for interaction control in human-human dialogue. Ultimately, software for improving interaction control in practical applications should capture all relevant cues.

The phenomena associated with turn-yielding include silent pauses, falling and rising intonation patterns, and certain vocal tract configurations such as exhalations (e.g. Duncan, 1972; Ford & Thompson, 1996; Local, Kelly & Wells, 1986). Turn-yielding cues are typically located somewhere towards the end of the contribution, although not necessarily on the final syllable. Granted that human turn-taking involves decisions above a reflex level, evidence suggests that turn-yielding cues must occur at least 200-300 ms before the onset of the next contribution (Ward, 2006; Wesseling & van Son, 2005).

The phenomena associated with turn-keeping include level intonation patterns, vocal tract configurations such as glottal or vocal tract stops without audible release, as well as a different quality of silent pauses as a result of these vocal tract closures (e.g. Caspers, 2003; Duncan, 1972; Local & Kelly, 1986). Turn-keeping cues are also located near the end of the contribution. As these cues are not intended to trigger a response, but rather to inhibit one, they may conceivably occur later than turn-yielding cues.

There are also a number of cues (in addition to the silent pauses mentioned above) that have been observed to occur with turn-yielding as well as with turn-keeping. Examples of such cues include decreasing speaking rate and other lengthening patterns towards the end of contributions. The mere presence (or absence) of such cues cannot be used for making a turn-yielding vs. turn-keeping distinction, although the amount of final lengthening, for example, might provide valuable guidance for such a task (cf. Heldner & Megyesi, 2003).

3 Prosodic cues applied to interaction control

In previous work (Edlund & Heldner, 2005), we explored to what extent the prosodic features extracted with **/nailon/** (Edlund & Heldner, forthcoming) could be used to mimic the interaction control behaviour in conversations among humans. Specifically, we analysed one of the interlocutors in order to predict the interaction control decisions made by the other person taking part in the conversation. These predictions were evaluated with respect to whether there was a speaker change or not at that point in the conversation, that is, with respect to what the interlocutors actually did.

Each unit ending in a silent pause in the speech of the interlocutor being analysed was classified into one out of three categories: turn-keeping, turn-yielding, and don't know. Units with low patterns were classified as suitable places for turn-taking (i.e. turn-yielding); mid and level patterns were classified as unsuitable places (i.e. turn-keeping); all other patterns, including high or rising, ended up in the garbage category don't know. This tentative classification scheme was based on observations reported in the literature (e.g. Caspers, 2003; Thórisson, 2002; Ward & Tsukahara, 2000), but it was in no way optimised or adapted to suit the speech material used.

This experiment showed that interaction control based on extracted features avoided 84% of the places where a system using silence duration thresholds only would have interrupted its users, while still recognizing 40% of the places where it was suitable to say something (cf. Edlund & Heldner, 2005). Interaction control decisions using prosodic information can furthermore be made considerably faster than in silence only systems. The decisions reported here were made after a 300-ms silence to be compared with silences ranging from 500 to 2000 ms in typical silence only systems (Ferrer, Shriberg & Stolcke, 2002).

4 Discussion

In this paper, we have discussed a number of prosodic features liable to be relevant for interaction control. We have shown that automatically extracted prosodic information can be used to improve the interaction control in spoken human-computer dialogue compared to systems relying exclusively on silence duration thresholds.

Future work will include further development of the automatic extraction in terms of improving existing algorithms as well as adding new prosodic features. In a long-term perspective, we would want to combine prosodic information with other sources of information, such as semantic completeness and visual interaction control cues, as well as to relate interaction control to other conversation phenomena such as grounding, error handling, and initiative.

Acknowledgements

This work was carried out within the CHIL project. CHIL (Computers in the Human Interaction Loop) is an Integrated Project under the European Commission's Sixth Framework Program (IP-506909).

References

Bell, L., J. Boye & J. Gustafson, 2001. Real-time handling of fragmented utterances. *Proceedings NAACL Workshop on Adaptation in Dialogue Systems*, Carnegie Mellon University, Pittsburgh, PA, 2-8.

Caspers, J., 2003. Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics 31*, 251-276.

Duncan, S., Jr., 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology 23*(2), 283-292.

Edlund, J. & M. Heldner, 2005. Exploring Prosody in Interaction Control. *Phonetica 62*(2-4), 215-226.

Edlund, J. & M. Heldner, forthcoming. /nailon/ – a tool for online analysis of prosody. *Proceedings 9th International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh, PA.

Ferrer, L., E. Shriberg & A. Stolcke, 2002. Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody in human-computer dialog. *Proceedings ICSLP 2002*, Denver, Vol. 3, 2061-2064.

Ford, C.E. & S.A. Thompson, 1996. Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E.A. Schegloff & S.A. Thompson (eds.), *Interaction and grammar*. Cambridge: Cambridge University Press, 134-184.

Heldner, M. & B. Megyesi, 2003. Exploring the prosody-syntax interface in conversations. *Proceedings ICPhS 2003*, Barcelona, 2501-2504.

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

Local, J.K., Kelly, J. & W.H.G. Wells, 1986. Towards a phonology of conversation: turn-taking in Tyneside English. *Journal of Linguistics 22*(2), 411-437.

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

Skantze, G. & J. Edlund, 2004. Robust interpretation in the Higgins spoken dialogue system. *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, Norwich.

Thórisson, K.R., 2002. Natural turn-taking needs no manual: Computational theory and model, from perception to action. In B. Granström, D. House & I. Karlsson (eds.), *Multimodality in language and speech systems.* Dordrecht: Kluwer Academic Publishers, 173-207.

Ward, N., 2006. Methods for discovering prosodic cues to turn-taking. *Proceedings Speech Prosody 2006*, Dresden.

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

Wesseling, W. & R.J.J.H. van Son, 2005. Early preparation of experimentally elicited minimal responses. *Proceedings Sixth SIGdial Workshop on Discourse and Dialogue*, ISCA, Lisbon, 11-18.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 57–60

SMTC – A Swedish Map Task Corpus

Pétur Helgason
Department of Linguistics and Philology, Uppsala University
petur.helgason@lingfil.uu.se

Abstract

A small database of high quality recordings of 4 speakers of Central Standard Swedish is being made available to the speech research community under the heading Swedish Map Task Corpus *(SMTC). The speech is unscripted and consists mostly of conversations elicited through map tasks. In total, the database contains approximately 50 minutes of word-labelled conversations, comprising nearly 8000 words. The material was recorded at the Stockholm University Phonetics Lab. This paper describes the recording method, the data elicitation procedures and the speakers recruited for the recordings. The data will be made available on-line to researchers who put in a request with the author (cf. section 7 below).*

1 Introduction

The data being made available under the heading *Swedish Map Task Corpus* (SMTC) were originally recorded as part of the author's doctoral dissertation project (Helgason, 2002). The data have already proved useful for several other research projects, e.g. Megyesi (2002), Megyesi & Gustafson-Čapková (2002) and Edlund & Heldner (2005). As it seems likely that future projects shall want to make use of the data, and the data are not described in much detail elsewhere, an account of the recording procedure and elicitation method are called for. At the same time, the data shall be made available for download for researchers.

2 Recording set-up

The data were recorded in the anechoic room at the Stockholm University Phonetics Lab. The subjects were placed facing away from one another at opposite corners of the room (see Figure 1). The "head-to-head" distance between the subjects was approximately two meters. The reason for this placement of the subjects was partly to minimize cross-channel interference, and partly to prevent them from consulting one another's maps (see the following section). The recording set-up was therefore in accordance with the nature of the data elicitation method.

The data were recorded using a Technics SV 260 A DAT recorder and two Sennheiser MKE2 microphones. Each microphone was mounted on a headset and placed in such a way that it extended approximately 2.5 cm out and to the side of the corner of the

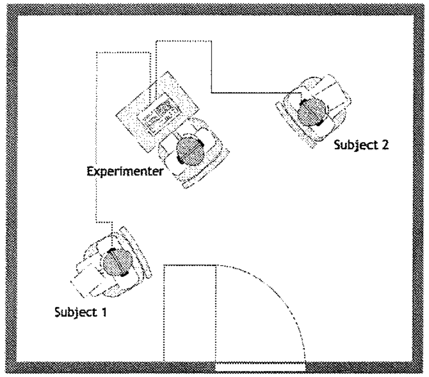

Figure 1. The placement of subjects and experimenter during the recording.