PROSODIC FEATURES IN THE VICINITY OF SILENCES AND OVERLAPS

Mattias Heldner
Jens Edlund
Kornel Laskowski
Antoine Pelcé

1 Introduction

In this study, we describe the range of prosodic variation observed in two types of dialogue context, using fully automatic methods. The first type of context is that of *speaker-changes*, or transitions from only one participant speaking to only the other, involving either acoustic silence or acoustic overlap. The second type of context is comprised of mutual silence or overlap where a speaker change could in principle occur but does not. For lack of a better term, we will refer to these contexts as *non-speaker-changes*. More specifically, we investigate F0 patterns in the intervals immediately preceding overlaps and silences – in order to assess whether prosody before overlaps or silences may invite or inhibit speaker change.

Previous work indicates that a number of prosodic and phonetic features are associated with speaker-changes and non-speaker-changes. With respect to F0 patterns, several studies have suggested that rising as well as falling pitch patterns are correlates of speaker-changes (e.g. Local & Kelly, 1986; Local, Kelly, & Wells, 1986; Ogden, 2001), and similarly that flat F0 patterns in the middle of a speaker's pitch range are correlates of non-speaker-changes (e.g. Caspers, 2003; Duncan, 1972; Koiso, Horiuchi, Tutiya, Ichikawa, & Den, 1998; Local & Kelly, 1986; Ogden, 2001; Selting, 1996). Furthermore, stretches of low F0 have been reported to invite backchannels in overlap as well as following silence (Ward & Tsukahara, 2000); flat intonation has also been reported to act as an inhibitory cue for backchannels (Noguchi & Den, 1998).

A fundamental problem in exploring prosody in dialogue lies in identifying locations at which prosody may turn out to be salient, and much of prior work has relied on the concepts of turns and floors, and thereby on manual or sufficiently accurate automatic detection of punctuation, disfluencies, and dialog act types. Frequently in naturally-occurring dialogue, these concepts are ill-defined. In previous work of our own, we investigated to what extent speaker-changes and non-speaker-changes can be predicted from a very limited number of F0 pattern

types (Edlund & Heldner, 2005), as well as from a direct representation of F0 variation (Laskowski, Edlund, & Heldner, 2008a, 2008b; Laskowski, Wölfel, Heldner, & Edlund, 2008), at locations dictated by low-level characterizations of the interactive state of the dialogue. In the present study, we take one step back to instead describe the range of diversity in F0 patterns occurring immediately before mutual silences or intervals of overlapping speech. We operationalize the annotation of these transitions using a standard finite state automaton over joint speech activity states. We then extract pitch variation features for these transition types and construct descriptive models to characterize them. An important contribution of this work is the visualization of these models, yielding an end-to-end methodology for zero-manual-effort analysis of pitch variation, conditioned on interactive dialogue context.

2 Methods

2.1 Materials

We used speech material from the Swedish Map Task Corpus (Helgason, 2006), designed as a Swedish counterpart to the HCRC Map Task Corpus (Anderson, et al., 1991). The map task is a cooperative task involving two speakers, intended to elicit natural spontaneous dialogues. Each of two speakers has one map which the other speaker cannot see. One of the speakers, the *instruction giver (g)*, has a route marked on his or her map. The other speaker, the *instruction follower (f)*, has no such route. The two maps are not identical and the subjects are explicitly told that the maps differ, but not how. The task is to reproduce the giver's route on the follower's map ("The design of the HCRC Map Task Corpus," n.d.).

Eight speakers, five females and three males, are represented in the corpus. The speakers formed four pairs, three female-male pairs and one female-female pair. Each speaker acted as instruction giver and follower at least once, and no speaker occurred in more than one pair. The corpus includes ten such dialogues, the total duration of which is approximately 2 hours and 18 minutes. The dialogues were recorded in an anechoic room, using close-talking microphones, with the subjects facing away from each other, and with acceptable acoustic separation of the speaker channels.

2.2 Procedures

The procedures involved defining, identifying and classifying instances of the two context types, extracting F0 patterns immediately before these, and summarizing and visualizing them. In this section, we outline and motivate how this was done.

2.2.1 Identifying interaction state transitions

As mentioned in the introduction, naturally occurring human-human dialogue contains a significant number of phenomena, such as backchannels, disfluencies, and cross-channel disruptions, which make it difficult to condition prosodic extraction on objectively defined syntactic or semantic boundaries. To address this problem, we limit ourselves to boundaries in conversation flow, defined by the relative timing of talkspurt deployment by the two parties. We annotate every instant in a dialogue with an explicit interaction state label; states describe the joint vocal activity of both speakers, building on a tradition of computational models of interaction (e.g. Brady, 1968; Dabbs & Ruback, 1984; Jaffe & Feldstein, 1970; Norwine & Murphy, 1938; Sellen, 1995). We note that, importantly, each participant's vocal activity is a binary variable, such that for example backchannel speech (Yngve, 1970) is not treated differently from other speech. We use the resulting conversation state labels to identify state transitions which define the end of the target intervals at which we subsequently extract prosodic features. The procedure involves three steps, as depicted in Figure 1.

First, we perform vocal activity detection, individually for each speaker, using the VADER voice activity detector from the CMU Sphinx Project ("The CMU Sphinx Group Open Source Speech Recognition Engines," n.d.). This results in the labeling of each instant, for each speaker, as either SPEECH or SILENCE.

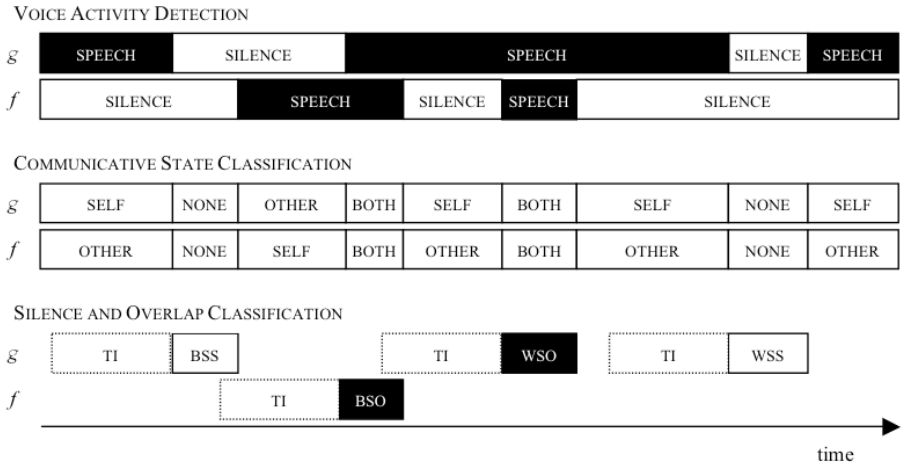

Figure 1. Illustration of how between-speaker silences (BSS), between-speaker overlaps (BSO), within-speaker silences (WSS), and within-speaker overlaps (WSO) are defined and classified, as well as how the target intervals (TI) are located with respect to these. The illustration shows all three steps (as in the text) from the perspectives of both *g* and *f*.

Second, at each instant, the states of the two speakers are combined to derive a four-class label of the communicative state of the conversation, describing both speakers' activity, from the point of view of each speaker. The four states we consider include SELF, OTHER, NONE and BOTH. For example, from the point of

view of the instruction giver *g*, the state is SELF if *g* is speaking and the instruction follower *f* is not; it is OTHER if *g* is silent and *f* is speaking, NONE if neither speaker is speaking, and BOTH if both are. The process of defining communicative states from the point of view of speaker *f* is similar; we illustrate this process for both speakers in the middle panel of Figure 1.

Finally, in a third step (comprising a third pass of the data, for illustration purposes), the NONE and BOTH states from Step 2 are further classified in terms of whether they are within- or between-speaker events, from the point of view of each speaker. This division leads to four context types: within-speaker overlap, SELF–BOTH–SELF; between-speaker overlap, SELF–BOTH–OTHER; within-speaker silence, SELF–NONE–SELF; and between-speaker-silence, SELF–NONE–OTHER. Speaker changes with neither overlap nor silence (i.e. with silence or overlap smaller than 10ms) are exceedingly rare in the material, and are not reported here. For completion, we note that the four states, per each of two speakers, together with the two states in which either *g* or *f* are speaking alone, constitute a 10-state finite state automaton (FSA) describing the evolution of dialogue in which only one party at a time may change vocal activity state. The number of states in such an interaction FSA may be augmented to model other subclassifications, or to model sojourn times, without loss of generality; here, we limit ourselves to an FSA of 10 states, and specifically to the 4 phenomena mentioned, as it is most directly relevant to our ongoing work in conversational spoken dialogue systems.

2.2.2 Extracting F0 patterns

Once the silences and overlaps are identified and classified, we collect F0 patterns from the last 500ms of speech in SELF-state preceding BSS, BSO, WSS and WSS (see the target intervals in Figure 1). It is in these intervals, approximately the last two syllables, *before* silences or overlaps, that we look for potential prosodic features inviting or inhibiting speaker-changes. The prosodic features we explored are all related to F0 patterns, but we use two different ways of capturing such patterns: one based on regular F0 extraction, and the other on a direct representation of F0 variation, known as the fundamental frequency variation spectrum.

The F0 estimates are computed using YIN (de Cheveigné & Kawahara, 2002). They are then transformed from Hertz to semitones, to make the pitch excursions of men and women more comparable. The data is subsequently smoothed using a median filter (over 9 10ms frames) to eliminate outlier errors. The resulting contours of smoothed F0 estimates are shifted along the vertical octave axis such that the median of the first three voiced frames in each contour falls on the mid-point of the y-axis. By plotting the contours with partially transparent dots, the visualizations give an indication of the distribution of different patterns with darker bands for concentrations of patterns and vice versa. We refer to this visualization as bitmap clustering.

In addition, we use a recently introduced vector-valued spectral representation of F0 variation – the fundamental frequency variation (FFV) spectrum – to capture F0 variation patterns (Laskowski, Edlund, et al., 2008a, 2008b; Laskowski, Wölfel, et al., 2008). Briefly, this technique involves passing the sequence of FFV spectra (a sample spectrum is shown in the left panel of Figure 2) through a filterbank (shown in the right panel of Figure 2), and inferring a statistical model over the filterbank representation.

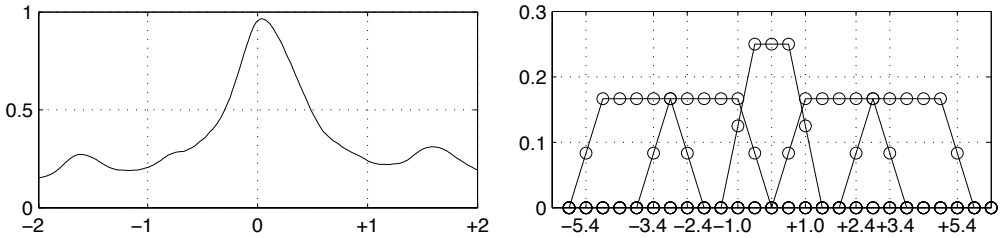

Figure 2. A sample fundamental frequency variation spectrum (left); the x-axis is in octaves per 8ms. Filters in the filterbank (right); the two extremity filters are not shown.

The filterbank attempts to capture meaningful prosodic variation, and contains a conservative filter for perceptually "flat" pitch, two filters for "slowly changing" rising and falling pitch, two filters for "rapidly changing" rising and falling pitch, and two wide extremity filters to capture unvoiced frames.

3 Results and discussion

From informal listening to the extracted regions, we observed that the instruction giver *g* and instruction follower *f* roles in the Swedish Map Task Corpus were somewhat unbalanced with respect to the kind of utterance types that occurred (see Cathcart, Carletta, & Klein, 2003 for similar observations in the HCRC Map Task Corpus). For example, whereas the speech before silences in the giver channel included a relatively high proportion of propositional statements, the follower channel instead contained a large proportion of continuers, that is backchannels indicating that the giver should go on talking (e.g. Jurafsky, Shriberg, Fox, & Curl, 1998) such as "mm" or "aa". Because of this imbalance, we decided to analyze giver prosody and follower prosody separately.

Table 1 shows the number of instances of interaction state transition types under study, given our definitions in Section 2.2.1. We note that, interestingly, the number of observed between-speaker phenomena, including silences and overlaps, is split evenly between the giver and follower roles, while the indications of imbalance with respect to roles is evident already in the relative proportions of the within-speaker phenomena.

Table 1. The number of observed interaction state transitions under study; relative proportion per speaker role shown is in parentheses.

Phenomenon	Giver		Follower		Total
Within-speaker silence (WSS)	1184	(78%)	334	(22%)	1518
Between-speaker silence (BSS)	977	(50%)	978	(50%)	1955
Within-speaker overlap (WSO)	276	(69%)	123	(31%)	399
Between-speaker overlap (BSO)	353	(51%)	346	(49%)	699

3.1 F0 patterns before between- and within-speaker silences (BSS & WSS)

Figure 3 shows bitmap cluster plots of F0 patterns during the 500ms preceding between- and within-speaker silences in the giver and follower channels. Our expectations before between-speaker silences included rising as well as falling F0 contours. As can be seen, there are falls and rises both in the giver and in the follower plots; broadly, the observations are in line with our expectations. However, it appears that there are relatively more falls in the giver plot, and relatively more rises in the follower plot. Furthermore, the falls tend to start earlier with respect to the subsequent silence than do the rises. These second-order trends are the subject of our ongoing exploratory analysis.

For the within-speaker silences, our expectations based on the literature were that we would observe mainly flat patterns. Indeed, in comparison to the between-speaker silences, there seem to be relatively fewer rises and falls and relatively more flat patterns in this context type. The plots for between-speaker silences have more of a fan or plume shape extending forward, whereas those for within-speaker silences are more tightly concentrated around the midline. We note that this concentration is to some extent an artifact of the shifting of the contours along the y-axis; the effect, however, is the same for all conditions.

3.2 F0 patterns before between- and within-speaker overlaps (BSO & WSO)

For between- and within-speaker overlaps, the expectations gathered from the literature were weaker. We are only aware of studies reporting that stretches of low pitch cue backchannels in overlap as well as in mutual silence (Ward & Tsukahara, 2000). We are not aware of any previous studies on F0 patterns in speaker-changes where one talkspurt, not explicitly categorized as a backchannel, is followed by another starting in overlap. From informal listening, we have observed that the between-speaker overlaps include backchannels (e.g. continuers) as well as utterances with more propositional content, and furthermore that the proportion of backchannels is higher in the within-speaker overlaps than in the between-speaker overlaps. Thus, to the extent that these instances include backchannels, low pitch patterns were expected.

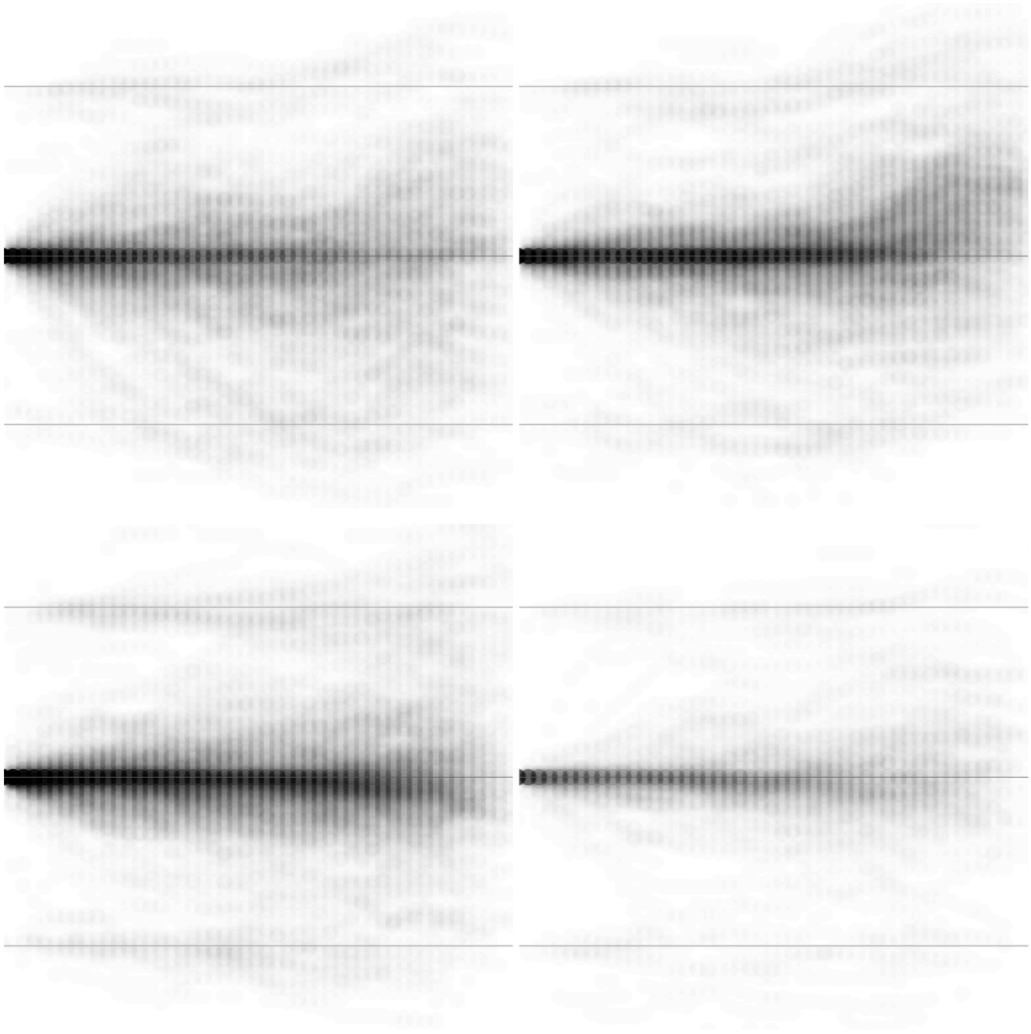

Figure 3. Bitmap clusters of TIs from *g* (left) and TIs from *f* (right) preceding between-speaker silences (top) and within-speaker silences (bottom). The y-axis represents F0 excursion from the median of the first three voiced frames in each 500ms TI in semitones. The thin grey horizontal lines indicate departures of an octave from the midline.

Our expectations were not borne out by the data, even after extending the analysis to the 1000 ms preceding the overlap; the authors of (Ward & Tsukahara, 2000) had found that stretches of low pitch were found at least 700ms before the onset of backchannels. Analysis of the corresponding graphical results (not shown due to space constraints) is the subject of our ongoing research. However, there is the possibility that speaker-change instances of overlap are governed by factors other than prosody, such as next speaker anticipating what current speaker will say. In these cases, next speaker may take the turn immediately, without either

heeding F0 signals or whether current speaker has finished or not – a situation which would explain the lack of consistent patterns in overlap conditions.

3.3 FFV spectra before between- and within-speaker silences (BSS & WSS)

The goal of the visualizations presented so far has been to describe extracted pitch contours preceding mutual silence and overlap. In this section, we investigate an alternative representation of those contours, namely hidden Markov models trained on the sequence of FFV spectra extracted from the same target speech intervals. As noted in (Laskowski, Edlund, et al., 2008b), the FFV representation offers several operational advantages over differencing F0 estimates.

Figure 4 shows two 4-state HMMs, for giver speech just prior to each of between-speaker silence and within-speaker silence. The emission probability of each state in both topologies is modeled using a single diagonal-covariance Gaussian. We note that the models describe the sequence of FFV spectra reversed in time, and, to retain nominal flow of time from left to right, entry into the model is shown from the right. Transitions whose probabilities are lower than 5% have been elided for clarity.

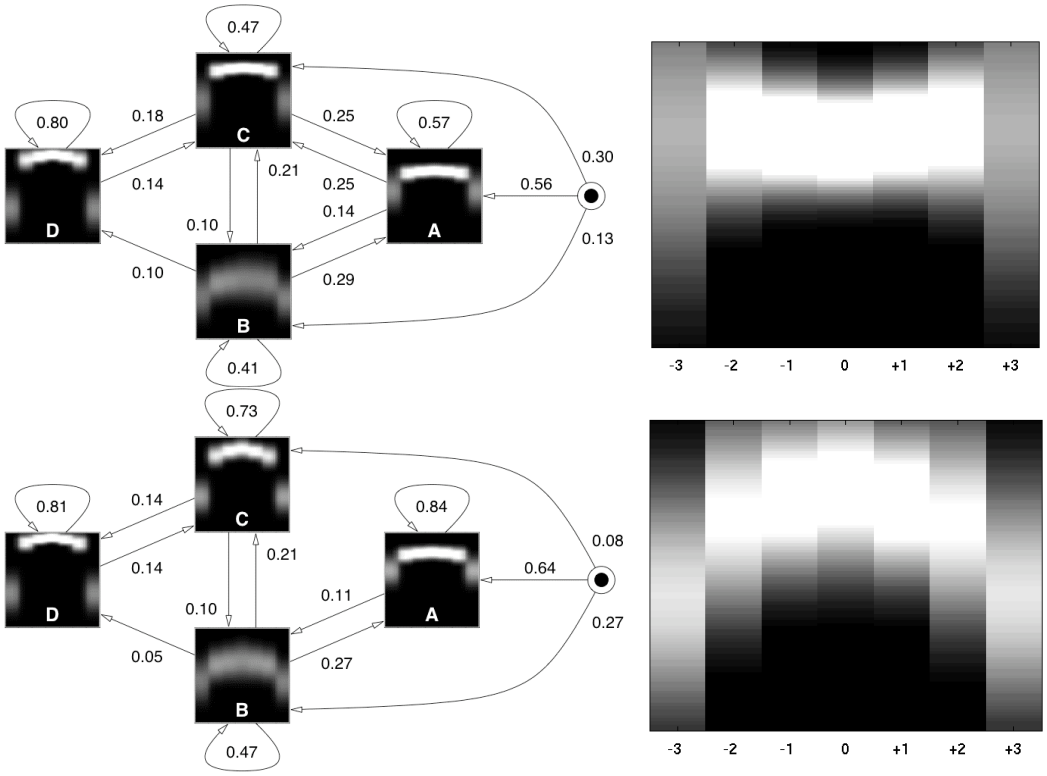

Figure 4**.** Inferred models for FFV spectrum sequences from giver speech preceding between- and within-speaker silences, at top and bottom, respectively. Filters are labeled from -3 to 3.

A comparison between the two topologies reveals that they are broadly similar, and that three of the states have similar emission probabilities. However, the states labeled "C" in both models differ in emission probability; to the right of each topology, we show a magnified view of that probability after the FFV spectra have been Z-normalized (as is performed during processing). It is evident that for the between-speaker model (top), the response of the 5 central filters (-2, -1, 0, 1, 2) is approximately equal, indicating a lack of preference for one of flat, slowly falling or rising, or quickly falling or rising pitch. In the within-speaker model (bottom), the central filter (0) corresponding to flat pitch contours shows a much higher response.

We note that the HMMs shown in Figure 4 have been successfully used to predict speaker changes (from *g* to *f*) in Swedish Map Task dialogues not used during model training (Laskowski, Edlund, et al., 2008a, 2008b; Laskowski, Wölfel, et al., 2008).

In Figure 5 we show a similar analysis for follower speech preceding silence. As in Figure 4, between-speaker silence is shown at the top while within-speaker silence is shown at the bottom.

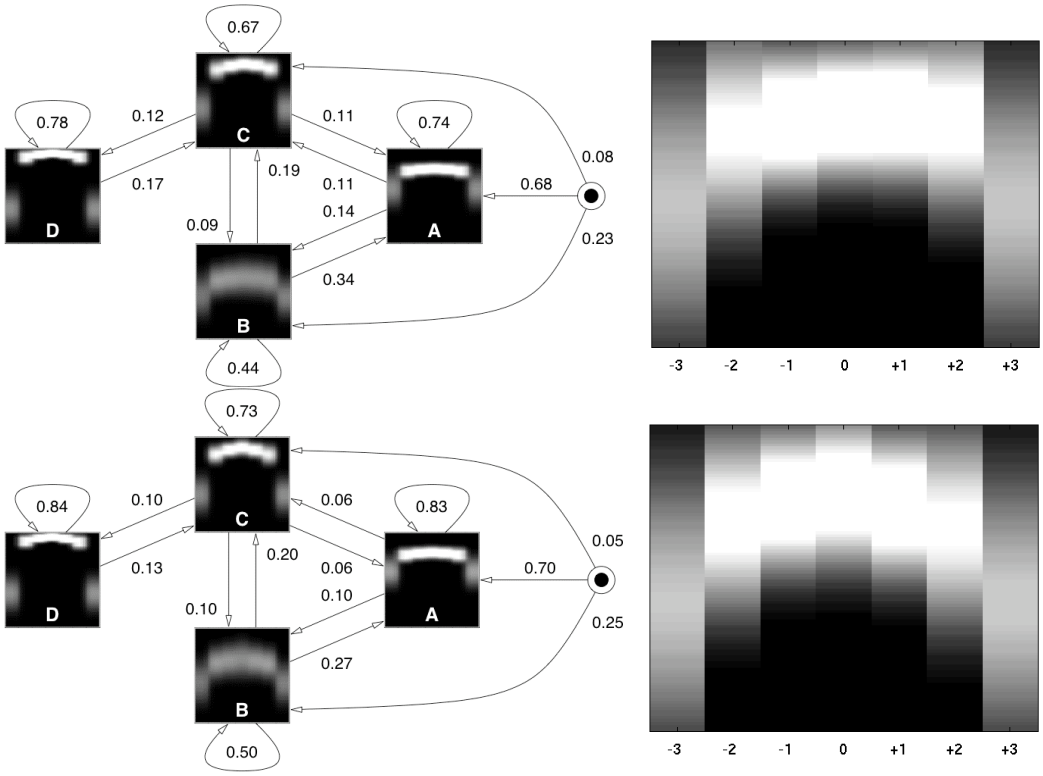

Figure 5**.** Inferred models for FFV spectrum sequences from follower speech preceding between- and within-speaker silence, at top and bottom, respectively. Filters are labeled from -3 to 3.

The two topologies are as similar as for instruction giver, but again there is a difference between the emission probabilities in the states labeled "C". When Z-normalized and magnified, the state in the within-speaker silence model (Figure 5, at bottom) shows a filterbank response which is visually similar to that for the same phenomenon in Figure 4. However, state "C" in the between-speaker silence model (top) shows stronger response for the filters corresponding to slowly and quickly rising pitch (filters 1 and 2) than for slowly and quickly falling pitch. In this regard, instruction follower's behavior deviates from that of instruction giver, suggesting that followers invite speaker change through rising pitch contours.

Conclusions

We have proposed and investigated an automatic methodology for the explorative study of prosodic features, which can inform spoken dialogue system design as well as generate hypotheses for testing in perceptual and pragmatic experiments. Furthermore, the methodology as proposed can form the starting point for further development of intonation models or phonological descriptions. The specific contributions of this effort are interconnected, and consist of the following.

1. We have defined in detail a zero-manual-effort means of locating potentially salient instants in dialogue, which relies on finite state automaton modeling of the joint vocal activity state of both dialogue participants. Locating these instants does not explicitly depend on syntactic or semantic criteria. The employed FSA may be easily extended in complexity to account for further sub-classification.
2. We have described two methods for characterizing and visualizing pitch variation in talkspurts and/or talkspurt fragments terminating at instants located in (1) above, namely bitmap clustering of normalized F0 estimates and HMM inference over FFV spectra.
3. The use of techniques in (1) and (2) above reveals visually distinguishable F0 contours preceding between-speaker silences and within-speaker silences, for both dialogue roles studied. Importantly, we have replicated the findings that between-speaker silences are often preceded by rising and falling pitch, whereas within-speaker silences are often preceded by flat pitch. This offers preliminary validation of the proposed methodology, recommending it as a basis for the automatic discrimination of silence types and for human-readable prosodic description. Furthermore, we have shown that patterns observed in other languages also occur in Swedish, for which, with the exception of our own work, this has not been shown previously.
4. The methodology reveals interesting differences between the giver and follower roles before between-speaker silences, notably relatively more falls in the giver's speech and relatively more rises in the follower's speech. This constitutes a novel finding that to our knowledge has not been reported by

other authors. Together with the informal observation that follower speech contains more backchannel continuers and giver speech more propositional statements, it opens possibilities for future sub-classification of between-speaker silence.

5. The methodology does not reveal visually distinguishable F0 contours preceding between-speaker overlaps and within-speaker overlaps.

Acknowledgements

This work was performed in co-operation between KTH and Carnegie Mellon University; we would like to thank Tanja Schultz and Rolf Carlson for continuing encouragement of this collaboration. We thank Pétur Helgason for access to the Swedish Map Task Corpus. Funding was provided in part by the Swedish Research Council (VR) project 2006-2172 Vad gör tal till samtal?

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., et al. (1991). The HCRC Map Task Corpus. *Language and Speech, 34*(4), 83-97.

Brady, P. T. (1968). A statistical analysis of on-off patterns in 16 conversations. *The Bell System Technical Journal, 47*, 73-91.

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

Cathcart, N., Carletta, J., & Klein, E. (2003). A shallow model of backchannel continuers in spoken dialogue. In *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 51-58). Budapest, Hungary.

Dabbs, J. M., Jr., & Ruback, R. B. (1984). Vocal patterns in male and female groups. *Personality and Social Psychology Bulletin, 10*(4), 518-525.

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America, 111*(4), 1917-1930.

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., & Heldner, M. (2005). Exploring prosody in interaction control. *Phonetica, 62*(2-4), 215-226.

Helgason, P. (2006). SMTC - A Swedish Map Task Corpus. In *Working Papers 52: Proceedings from Fonetik 2006* (pp. 57-60). Lund, Sweden.

Jaffe, J., & Feldstein, S. (1970). *Rhythms of dialogue*. New York, NY, USA: Academic Press.

Jurafsky, D., Shriberg, E., Fox, B. A., & Curl, T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers* (pp. 114-120). Montreal, Canada.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech, 41*(3-4), 295-321.

Laskowski, K., Edlund, J., & Heldner, M. (2008a). An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems. In *Proceedings ICASSP 2008* (pp. 5041-5044). Las Vegas, NV, USA.

Laskowski, K., Edlund, J., & Heldner, M. (2008b). Machine learning of prosodic sequences using the fundamental frequency variation spectrum. In *Proceedings of the Speech Prosody 2008 Conference* (pp. 151-154). Campinas, Brazil: Editora RG/CNPq.

Laskowski, K., Wölfel, M., Heldner, M., & Edlund, J. (2008). Computing the fundamental frequency variation spectrum in conversational spoken dialogue systems. In *Proceedings of the 155th Meeting of the Acoustical Society of America, 5th EAA Forum Acusticum, and 9th SFA Congrés Français d'Acoustique (Acoustics2008)* (pp. 3305-3310). Paris, France.

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

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

Noguchi, H., & Den, Y. (1998). Prosody-based detection of the context of backchannel responses. In *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP'98)* (pp. 487-490). Sydney, Australia.

Norwine, A. C., & Murphy, O. J. (1938). Characteristic time intervals in telephonic conversation. *The Bell System Technical Journal, 17*, 281-291.

Ogden, R. (2001). Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association, 31*, 139-152.

Sellen, A. J. (1995). Remote conversations: The effects of mediating talk with technology. *Human-Computer Interaction, 10*, 401-444.

Selting, M. (1996). On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics, 6*, 357-388.

The CMU Sphinx Group Open Source Speech Recognition Engines (n.d.). from http://cmusphinx.sourceforge.net/

The design of the HCRC Map Task Corpus (n.d.). from http://www.hcrc.ed.ac.uk/maptask/maptask-description.html

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

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting Chicago Linguistic Society* (pp. 567-578). Chicago, IL, USA: Chicago Linguistic Society.