

Prosodic features associated with the distribution of turns in Finnish informal dialogues

Mietta Lennes

Hanna Anttila

Helsingin yliopisto, Fonetikan laitos
mietta.lennes@helsinki.fi

Helsingin yliopisto, Fonetikan laitos
hanna.anttila@helsinki.fi

Abstract

In free dialogue, the speaking time may be distributed among the speakers in various ways. These turn-taking dynamics probably reflect interactional settings. The timing of utterances and pauses has been studied as early as in the 1930's, whereas the long-term prosodic properties of interaction and turn-taking dynamics have received less attention. In this preliminary study, we explore and visualize fragments of Finnish informal dialogue with regard to turn-taking dynamics and such prosodic features as F0, speech rate, and creaky voice.

1 INTRODUCTION

In ordinary, casual dialogue, the speaking time may be distributed among the speakers in various ways. For a certain period of time, one of the speakers may tend to use long turns, while the other participant acknowledges these with short utterances. For another while, both speakers may keep taking either longer or shorter turns. Speakers may interrupt or overlap each other at certain points. During a conversation, these *turn-taking dynamics* may vary a great deal, and they probably reflect certain interactional settings (see, e.g., Hakulinen, 1997). It is possible that they are related to systematic changes in the prosodic and acoustic properties of speech, but such processes are not well understood.

Recently, the general interest for the phonetic research of free conversational speech has significantly increased. One attempt to emphasize this kind of research is the international project *Spontaneous Speech of Typologically Unrelated Languages (Russian, Finnish and Dutch): Comparison of Phonetic Properties* (INTAS 915), which was launched in 2001. This project has produced high-quality recordings of informal Finnish dialogues, which are currently being annotated. In the present article, we will both explore fragments of this material by visualization techniques and outline some of our preliminary observations regarding the turn-taking dynamics in these dialogues.

2 BACKGROUND

The long-term aspects of turn-taking in spoken interaction were considered as important objects of study as early as in the 1930's. Turn-taking habits were thought to reflect the speaker's personality, thus even providing a possible method for clinical assessment. This interest led to the development of a special instrument, the *interaction*

chronograph, by an anthropologist called Chapple and his colleagues (see Matarazzo et al, 1956). The device would draw continuous lines on a moving tape when the observer pressed down a key. This way, one could manually record the start and end points of utterances from the speakers in a live conversation. The corresponding field of study became known as *chronography*.

Chronographic analysis was facilitated when computer equipment became available to the scientific community. Jaffe and Feldstein (1970) discovered that the timing of utterances and pauses had certain mathematical properties that remained relatively constant over speakers, such as the distribution of utterances and pauses of different durations. Chronographic research was pursued also in Finland. In a computerized chronographic analysis of telephone conversations, Sneek (1987) found that Finns tended to have longer pauses in comparison to Americans, and that Finns also tolerated longer pauses. Interestingly, the language-related differences diminished in intercultural conversations, suggesting that speakers adapt to each other's chronography. Sneek also noted that the amount of simultaneous speech was usually very small (a few percent of the total conversation time).

Due to earlier technological restrictions, chronographic studies have only dealt with properties related to the timing of utterance starts and ends, i.e., overlaps and pauses. These are only a small part of spoken interaction. Since regions of different turn-taking dynamics can be found in spoken dialogue, and since they may reflect certain interactional settings, it is necessary to find out whether a combination of chronographic and long-term prosodic analyses could add to our understanding of turn-taking processes.

Any prosodic pattern may have different functions depending on its context. The prosodic phenomena associated with turn-taking are not well known. For instance, Ogden (2001) suggests that creaky voice and glottal stops may have different functions as turn-holding or turn-switching devices in Finnish talk. Schaffer (1983) found that rising intonation is the strongest perceptual cue for turn switching in English, but she noted that there is great variability in such cues.

Topic changes are also often associated with certain prosodic phenomena. This is probably most evident in read speaking styles (see, e.g., Tür, Hakkani-Tür, Stolcke, and Shriberg, 2001, on automatic topic segmentation from news broadcasts). In read speech, mere pause duration is a strong cue of topic shifts. Some studies have been performed on spontaneous speech with the conclusion that topic shifts are marked with similar cues as sentence or utterance boundaries (see, e.g., Swerts, 1997). However, most of these studies have dealt with monologue speech (e.g., instruction monologue), and few studies are available on spontaneous or informal dialogue or conversation, where quite different factors come into play. For instance, such terms as "paragraph intonation" are not suitable for describing free conversation. Schaffer (1984) found that the prosodic topic boundary cues that seem to exist in read speech are not necessarily applicable to free dialogue, which may be explained by the different length of turns in (read) monologue vs. dialogue. In free conversation, the syntactic and semantic information provides much stronger topic cues than intonation. She also noted that in casual dialogue speech, intonation may not have to signal topic changes as much, since both speakers are involved in the same topic, but instead it may be needed as a turn transition cue.

We suspect that people may also perceive long-term prosodic properties that extend over utterance boundaries, since any incoming speech is interpreted in relation to a larger context (be it linguistic, interactional, or phonetic). This kind of global prosodic changes require further phonetic research.

3 METHODS

Recordings of five informal dialogues were used as data. The speakers were ten young Finnish adults (aged 20-30 years; five females) from the capital city region in Finland. The participants of each dialogue were close friends.

The recordings were performed in an anechoic room. The speakers were sitting a couple of meters apart and they were facing opposite walls to prevent eye contact and gestural communication. Each speaker's voice was recorded with a head-mounted microphone on a separate channel of a DAT recorder. The speakers were instructed to talk freely with each other, and they were left alone in the room for a total of 40 to 50 minutes. The speakers were checked upon every ten minutes and given some topics to discuss in case they would run out of topics of their own.

The recorded material was transferred to a computer and the two channels were separated, producing one long sound file per speaker. The utterance boundaries for each speaker were marked and the utterances were orthographically labeled. Clearly distinguishable changes in topic were determined and marked on the basis of these transliterations. Creaky-voiced regions were marked as well. Using these annotations as landmarks, the desired parts of the dialogues could be automatically analyzed. The Praat program (Boersma) was used for segmentation, labeling, and the acoustic analyses.

4 ANALYSES AND RESULTS

Assuming that a lot of changes in turn-taking dynamics probably occur around major changes in topic, we selected excerpts around those time points that were marked as involving a topic change. These fragments were then explored by automatically measuring and visualizing their temporal structure, fundamental frequency structure, voice quality changes, and speech rate.

4.1 Distribution of speaking time

In the different parts of a dialogue, the dynamics of turn-taking can be radically different. Figure 1 illustrates the actual temporal distribution of utterances in two fragments within dialogue 5, recorded from two male speakers, M1 and M2. The duration of each fragment was 3:20 minutes (200 seconds).

Figure 2 represents the distribution of total speaking time for each of the different dialogues and speakers. Apparently, the amount of speaking time per speaker was fairly well balanced in each dialogue. The amount of silence also contains some intervals of speech that were excluded from further analysis for technical reasons, so this measure should in fact not be taken into account.

Overlapping speech refers to those parts of the participants' utterances that occur simultaneously in time. Overlapping intervals of a duration of 300-700 ms were most frequent in all dialogues, indicating that most overlaps are produced in the backchannel, as "minimal feedback". Similar results have been presented by Jaffe and Feldstein (1970): overlaps are typically very short in duration.

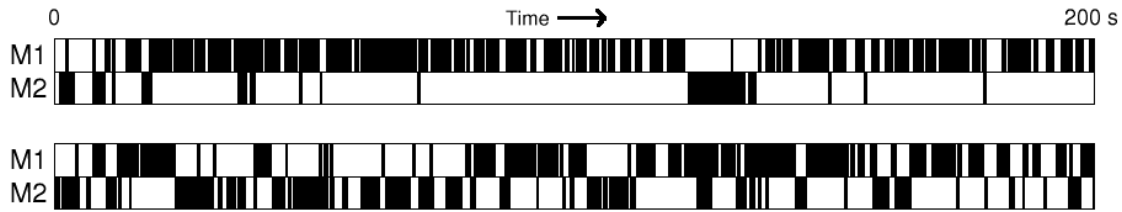


Figure 1. Two fragments exhibiting the variable turn-taking dynamics in dialogue 2 by two male speakers, M1 and M2. Black rectangles indicate the temporal intervals during which each speaker was speaking. In the topmost fragment, speaker M1 seems to dominate the interaction, whereas the second fragment shows more turn switches.

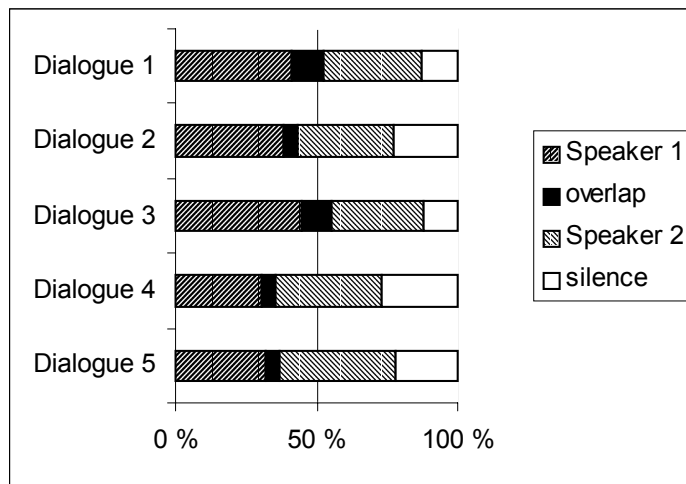


Figure 2. Overall distribution of speaking time in five dialogues.

4.2 Fundamental frequency

We wanted to describe how the fundamental frequency range of the speakers may vary in the different parts of a dialogue and how this range might relate to turn-taking dynamics. The definition of the fundamental frequency range of any particular speaker is problematic, and it is difficult to describe changes in F0 contours with statistical measures. Therefore, we decided to simply plot the fundamental frequency curves from a given set of utterances on top of each other. A similar procedure has been used by Iivonen (1999) in his Temporal Voice Range analysis. Our F0 curves are plotted as time-normalized to give an idea of the typical F0 contour of the utterances. The F0 points were calculated at 10 ms intervals.

In order to be accurate, the automatic algorithms for pitch extraction require that the upper and lower limits of the speaker's F0 range are given as parameters. Undesirable results may be caused by inappropriately defined F0 limits or by non-modal voice quality, such as creak, which is quite common in informal Finnish speech. We defined suitable F0 limits for each individual speaker by visually inspecting a number of F0 plots. Although fundamental frequency is related to the perceived pitch of speech, this relationship is nonlinear due to the properties of the human auditory system. We decided to plot all F0 curves using the semitone scale, which better corresponds to the perceived relative pitch changes. This also helped us to compare the different speakers irrespective of their individual differences in pitch range. All pitch curves are scaled as 20 semitones upwards from the speaker's lower F0 limit.

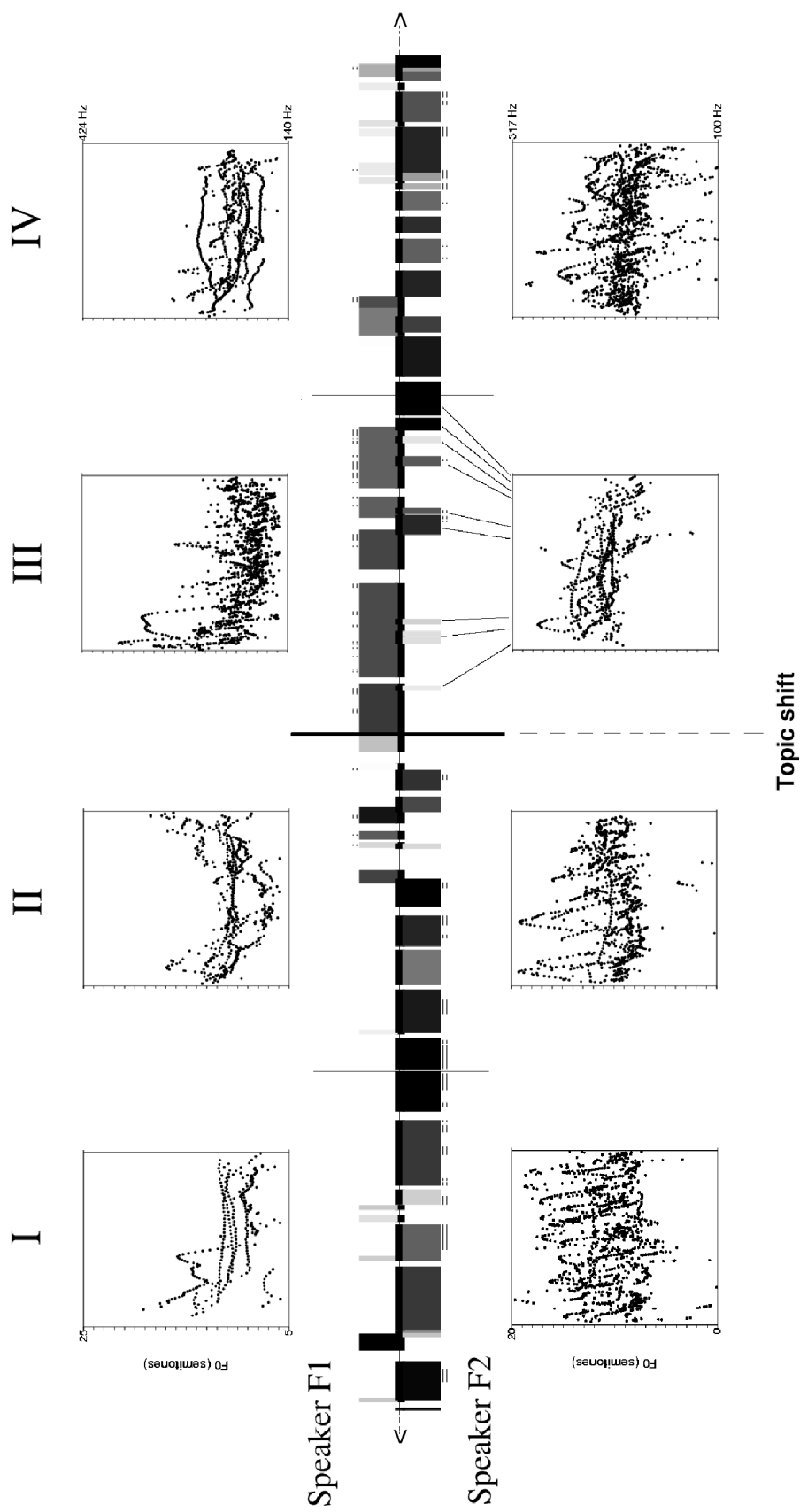


Figure 3. Utterances of two female speakers within a two-minute dialogue fragment are shown as shaded rectangles. A change in topic occurs in the middle of the fragment. Speaking rate (phonemes/second) is indicated by shades of gray in the rectangles: darker shade refers to faster speech. Creaky-voiced sections are shown as double lines. The time-normalized fundamental frequency curves are plotted in semitone scale for all the utterances within each of the four time regions. Note how the floor is switched from speaker F2 to F1 and then returned to F2.

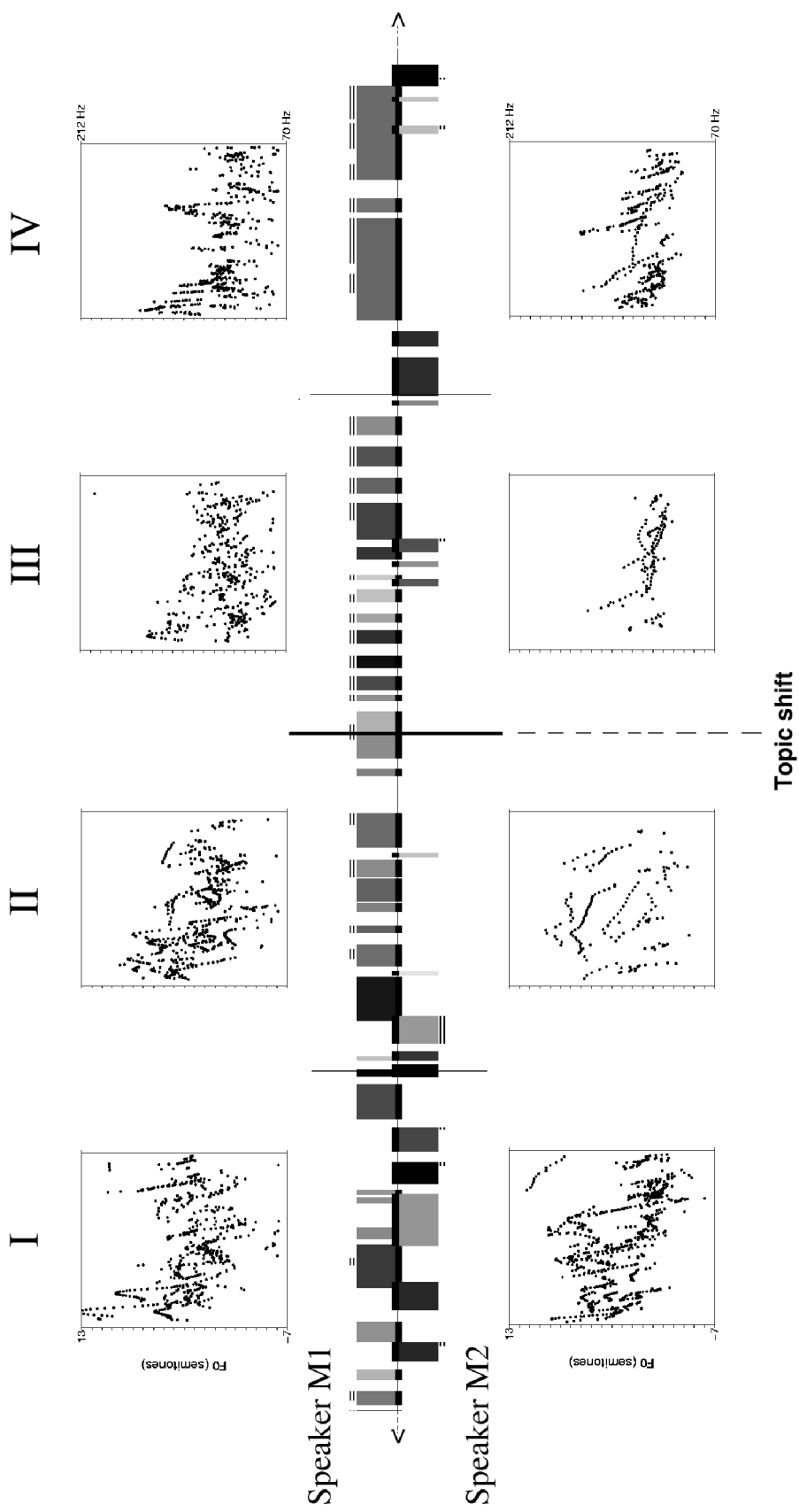


Figure 4. Utterances of two male speakers within a two-minute dialogue fragment are shown as shaded rectangles. A change in topic occurs in the middle of the fragment. Speaking rate (phonemes/second) is indicated by shades of gray in the rectangles: darker shade refers to faster speech. Creaky-voiced sections are shown as double lines. The time-normalized fundamental frequency curves are plotted in semitone scale for all the utterances within each of the four time regions. From region II onwards, the speaker M2 is mostly in the backchannel.

Figures 3 and 4 show two exemplary dialogue fragments, which were selected around a topic change or shift. The fragment in figure 3 involves two female speakers, F1 and F2 (dialogue 1), who are talking about school. F2 is first telling about her grades and says she was not very good at mathematics (the leftmost regions I and II). She jokes about being capable of checking her own salary or counting a reduced price in a shop. Speaker F1 agrees and comes to think of high schools that have specialized on certain fields, such as mathematics (this is where the topic shift occurs). Speaker F2 accepts the new topic and talks about high schools specialized in sports and golf. The pitch contours for the utterances of each of the four regions show that the general shape of pitch contours changes in the different parts of the fragment. If a speaker is “in the backchannel”, producing short utterances only (*mm*, *joo*, *nii*), the pitch contours are usually few, and often rather flat or slightly lowering. When the speaker “has the floor” and is talking vividly, the pitch range grows larger. Emotional factors probably contribute to pitch range as well. In read speech (e.g. newsreading), a high initial pitch is often used to mark a topic change. A similar phenomenon seems to occur in the dialogue fragment.

Figure 4 shows a fragment from dialogue 2 between two male speakers M1 and M2, who are also talking about school. The topic switch occurs after M1 has been telling about his class and when he starts to talk about a school trip. During the fragment, the speaker M2 switches into the backchannel to attend to M1’s story. In this fragment, several rhythmically different regions are visible. However, the topic shift is handled within a single long turn of the speaker M1, and a high initial pitch at the beginning of a new topic is not as evident as in figure 3.

4.3 Tempo

Since our material had not yet been fully segmented in the level of speech sounds or syllables, we were not able to use fine-grained, time-sensitive measures for speech rate or tempo. However, since the general idea in this study was to view a dialogue from a more global angle, we decided to settle for a rough definition of tempo: the number of characters of the orthographic transliteration of each utterance was divided by the duration of the utterance. Since the Finnish orthographic system is said to exhibit a nearly one-to-one correspondence between graphemes and phonemes, and since phonemic length is generally marked by doubling the corresponding written character, it may be claimed that this measure can give a rough description of speaking tempo as “phonemes” per second in the utterance level. These speech rate values are shown in figures 3 and 4 as shades of gray: darker shade refers to higher speech rate. The highest rates tend to occur around turn switches or in the middle of longer turns. However, in making detailed conclusions on the basis of our rough speaking rate measure, caution is well in order, and the results need to be further investigated.

Our speech rate measure is somewhat distorted in the case of very short utterances (e.g., *mm*, *joo*, *nii*), where slight errors in the utterance boundary placement have a strong effect, and the common [h]-like (or weakly voiced) “tail” of those utterances increases the mean phonemic duration. Seven speakers out of ten had the type of bipolar distribution seen for speaker F1 in figure 5. These speakers tended to use short “feedback” words (such as *joo* or *nii*) in the backchannel, whereas for instance speaker M2 (see figure 5) avoided these words.

In addition, we calculated the mean speech rates for each speaker as the mean number of words per second. This speed varied across speakers from 2,85 to 3,67 words per second. Interestingly, the participants in each dialogue had often very similar mean speech rates, suggesting that some adaptation or imitation may occur.

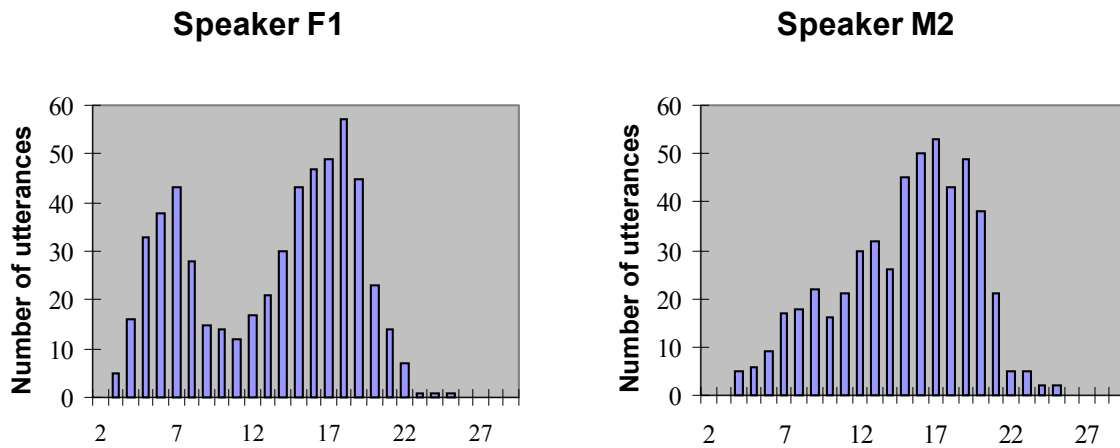


Figure 5. Distribution of utterances with different speech rates from a female speaker F1 and a male speaker M2. Speech rate (horizontal axis) was measured as the number of phonemes per second within each utterance.

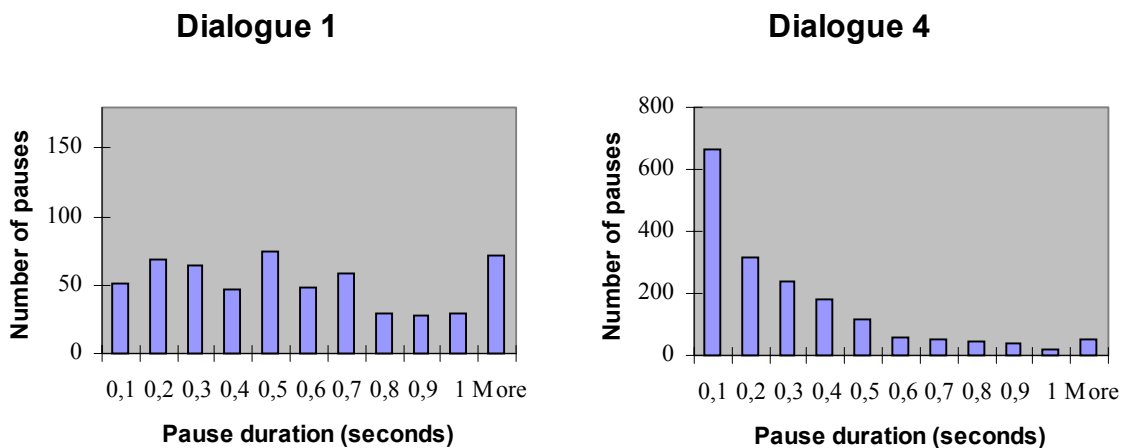


Figure 6. The distributions of pauses of different durations for two dialogues. The exponential-like pause distribution in Dialogue 4 is probably more accurate, since the utterance boundaries were more carefully marked.

4.4 Pauses

In this study, pauses were implicitly defined by utterance boundaries. Thus, a pause is the region between two utterances, during which the speaker is not articulating. Different pause types were not distinguished. For instance, in this definition pauses do not include the phonetic phenomena that occur at the boundary of intonation phrases or intonation units, unless these units are separated by a silence or a “non-articulatory” sound (hesitation creak or breathing noise). Figure 6 shows the distribution of the pauses of different durations. The distribution of pause durations in dialogue 4 closely resembles an exponential distribution, which is due to the fact that the utterance boundaries in this dialogue were much more accurately segmented than those in the other dialogues (e.g., dialogue 1 in figure 6). An exponential distribution of pause durations as well as utterance durations was found to be typical by Jaffe and Feldstein

(1970). Pauses longer than approximately 200 msec are often considered as signals of some kind of discontinuity in the interaction, which may be associated with their less frequent occurrence. It should be noted that since segmentation accuracy has such a strong effect on the distribution of pause durations, detailed conclusions cannot be made unless the utterance boundaries have been carefully marked.

4.5 Creak

All ten speakers used a creaky voice quality to some extent. A comparison of the distribution of creak within utterances showed that creak can occur at any point within an utterance, but it is more common towards the end. However, there are individual differences in the way creak is used, and we were not able to determine what kind of relationships there might be between creak, turn transitions and more global turn-taking dynamics.

5 CONCLUSIONS

In natural dialogue, many social, semantic, pragmatic, linguistic, and phonetic phenomena are intertwined. Studies on the different properties of dialogue speech generally deal with written transcripts and/or acoustic analyses of specific utterances in context. However, any live conversation is subject to constant change and adaptation. The long-term variations in the rhythm of the conversation are not easily illustrated in written transcripts, and they may not be visible in the acoustic analyses of only a few seconds of speech.

Using different visualization techniques in combination with acoustic measurements and linguistic information, it is possible to explore dialogue speech from a more global angle, which can give rise to interesting and valuable new hypotheses concerning interactional and phonetic processes. The acoustic measures used in the present study are still rough, and careful research is needed to show which kind of measures best reflect the long-term properties of turn-taking dynamics. Combining these objective acoustic-phonetic measures with an interactional analysis could help us to better understand the dynamics of dialogue.

6 ACKNOWLEDGMENTS

This study was supported by INTAS (project nr. 915) and the Academy of Finland (project *Integrated resources for speech technology and spoken language research in Finland*).

REFERENCES

- Boersma, P. and Weenink, D. 1992-2002. Praat – doing phonetics by computer. Available at: <http://www.praat.org/>.
- Jaffe, J. and Feldstein, S. 1970. *Rhythms of dialogue*. New York: Academic Press.
- Hakulinen, A. 1997. Vuorottelujäsennys. In: Tainio, L. (ed.) *Keskusteluanalyysin perusteet*. Tampere, Finland. Vastapaino, pp. 32–55.
- Iivonen, A. 1999. F0 contours of utterances superimposed on the temporal voice range profile of the speaker. *Proceedings of the XIV International Congress of Phonetic Sciences 1999*, San Francisco 1-7 Aug. , 953-956.

- Matarazzo, J., Saslow, G. and Matarazzo, R. 1956. The interaction chronograph as an instrument for objective measurement of interaction patterns during interviews. *Journal of Psychology* 41, 347-367.
- Ogden, R. 2001. Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association* 31, 139-152.
- Schaffer, D. 1983. The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics* 11, 243-257.
- Schaffer, D. 1984. The role of intonation as a cue to topic management in conversation. *Journal of Phonetics* 12, 327-344.
- Sneck, S. 1987. Assessment of chronography in Finnish-English telephone conversation: an attempt at a computer analysis. *Jyväskylä Cross-Language Studies* 14. Department of English, University of Jyväskylä.
- Swerts, M. 1997. Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America* 101, pp. 514-521.
- Tür, G, Hakkani-Tür, D, Stolcke, A. and Shriberg, E. 2001. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics* 27, pp. 31-57.

All Praat scripts that were used to analyse the data described in this paper will be made available at: <http://www.helsinki.fi/~lennes/praat-scripts/>