Journal of Phonetics 38 (2010) 555–568

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Pauses, gaps and overlaps in conversations

Mattias Heldner *, Jens Edlund

KTH Speech, Music and Hearing, Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden

ARTICLE INFO

Article history:
Received 18 December 2008
Received in revised form
18 December 2009
Accepted 3 August 2010

ABSTRACT

This paper explores durational aspects of pauses, gaps and overlaps in three different conversational corpora with a view to challenge claims about precision timing in turn-taking. Distributions of pause, gap and overlap durations in conversations are presented, and methodological issues regarding the statistical treatment of such distributions are discussed. The results are related to published minimal response times for spoken utterances and thresholds for detection of acoustic silences in speech. It is shown that turn-taking is generally less precise than is often claimed by researchers in the field of conversation analysis or interactional linguistics. These results are discussed in the light of their implications for models of timing in turn-taking, and for interaction control models in speech technology. In particular, it is argued that the proportion of speaker changes that could potentially be triggered by information immediately preceding the speaker change is large enough for reactive interaction controls models to be viable in speech technology.

1. Introduction

Within a larger project to investigate and model speech phenomena that are specific to conversations, and in particular prosodic aspects of the control of the interaction in conversation, this study explores certain durational aspects of spoken interaction. The central topic of this paper is between-speaker and within-speaker intervals in conversations. These intervals include acoustic silences in the conversation as well as stretches of overlapping speech. We examine the statistical distribution of intervals *within* the speech of one speaker (i.e. pauses) as well as *between* speakers or at speaker changes (i.e. gaps and overlaps). The interest in pause, gap and overlap distributions in conversations is motivated from several perspectives.

From a basic research perspective, it is compelling to collect empirical data that can be used to test theories and models of spoken interaction. For example, such data may potentially support or rule out models using the cessation of speech, the acoustic silences themselves, or prosodic cues towards the end of utterances as interaction control signals in conversations. Perhaps more importantly, the data can also be used to quantify the number of cases where it is possible that *reaction to* such signals is relevant to speaker changes. It is clear that the decision to start speaking in overlap with the previous speaker cannot be a reaction to the offset of speech, to silence or to some prosodic information immediately before the silence. Given what is known about minimal response times for spoken utterances (Fry, 1975; Izdebski & Shipp, 1978; Shipp, Izdebski, & Morrissey, 1984), however, there may be speaker changes where the gap is long enough for the next speaker to react to such information, and we would like to know how frequently this is the case.

From an applied research perspective – that of speech technology – pause, gap and overlap distributions in conversations are interesting, among other things, because they give an indication of what kind of interaction control behavior to aim for in conversational spoken dialogue systems. For example, we would like to know what range of response times such systems should be able to produce. Such timing is vital for systems with the ambition of presenting more human-like behavior (Edlund, Gustafson, Heldner, & Hjalmarsson, 2008; Gustafson, Heldner, & Edlund, 2008), or as Cassell puts it, for "a machine that acts human enough that we respond to it as we respond to another human" (Cassell, 2007, p. 350). The relation between pause and gap distributions, furthermore, has implications for the usefulness of silence duration thresholds for end-pointing in speech technology applications. Quantifying the proportion of speaker changes that can potentially be achieved by direct reaction to some interaction control signal is also of interest for speech technology, as it is foreseeable that spoken dialogue systems will have to resort to reactive methods with respect to turn-taking in the near future. This is a main concern in this paper: we want to corroborate the idea of using prosodic information for interaction control in speech technology applications.

In conversation analysis or interactional linguistics, finally, the oft-quoted claim that human turn-taking is so precise that next speakers tend to start with *no gap* and *no overlap* (Sacks, Schegloff,

* Corresponding author. Tel.: +46 8790 7563; fax: +46 8790 7854.
E-mail address: heldner@kth.se (M. Heldner).

0095-4470/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2010.08.002

& Jefferson, 1974), henceforth no-gap–no-overlap, is 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-endings, and furthermore that this projection is based solely on syntactic information (e.g. de Ruiter, Mitterer, & Enfield, 2006; Levinson, 1983). The between-speaker interval distributions presented in this paper provide empirical evidence that can support or challenge these claims.

A few methodological issues regarding the statistical treatment of durational data are brought to the surface by our goal of examining pause, gap and overlap distributions. Specifically, this concerns logarithmic transformation of durational data and whether to split the data into gaps and overlaps or to represent the intervals in speaker changes as one continuum.

This paper is composed as follows: we first give a background, including a survey of previous studies on pause, gap and overlap duration distributions. Next, we describe the methods, materials and procedures used. The results section first presents our findings regarding the statistical treatment of durational data. Then, we quantify the proportion of pauses, gaps, overlaps and no-gap–no-overlaps in genuine conversations. We present distribution analyses of the duration of pauses, gaps and overlaps extracted from three different datasets and for three different languages, and use these analyses, among other things, to quantify the proportion of speaker changes that could potentially be a reaction to the offset of speech or to some prosodic information immediately before the silence, as well as the ones that have to rely on other information. Finally, we discuss implications of these results for timing in turn-taking theories as well as for speech technology applications.

2. Background

2.1. Terminology for between- and within-speaker intervals

Silences and overlaps in conversations have received a lot of attention, and a large number of terms have been coined for very similar concepts, and especially so for silences at speaker changes. Sacks et al. (1974) distinguished between three kinds of acoustic silences in conversations: *pauses, gaps,* and *lapses.* This classification was based on what preceded and followed the silence in the conversation, and on the perceived length of the silence. Pauses, in this account, referred to silences within turns; gaps referred to shorter silences between turns or at possible completion points (i.e. at *transition-relevance places* or *TRPs*); and lapses referred to longer (or extended) silences between turns. However, the classification was complicated by the fact that the right context of the silence was also taken into account. For example, a silence followed by more speech by the same speaker would always be classified as a pause; also if it occurred at a TRP. Although this situation was not mentioned in the text, it seems fair to assume that any silence followed by a speaker change would be classified as a gap or a lapse also when it did not occur at a TRP. Hence, gaps and lapses could in practice only occur when there was a speaker change. There is also the possibility of speaker changes involving overlaps or no-gap–no-overlaps, which were the terms used by Sacks et al. (1974).

In addition to gaps, it seems that just about any three-way combination of (i) inter/between, (ii) turn/speaker, and (iii) silences/pauses/intervals/transitions have been used for concepts similar to gaps and duration of gaps at some point in time (e.g. Bull, 1996; Roberts, Francis, & Morgan, 2006; ten Bosch, Oostdijk, & Boves, 2005; ten Bosch, Oostdijk, & de Ruiter, 2004b). Other closely related terms include *(positive) response times* (Norwine & Murphy, 1938), *alternation silences* (Brady, 1968), *switching pauses* (Jaffe & Feldstein, 1970), *(positive) switch time* or *switch pauses* (Sellen, 1995), *transition pauses* (Walker & Trimboli, 1982), *(positive) floor transfer offsets* (de Ruiter et al., 2006), or just *silent or unfilled pauses* (e.g. Campione & Véronis, 2002; Duncan, 1972; Maclay & Osgood, 1959; McInnes & Attwater, 2004; Weilhammer & Rabold, 2003).

Pauses and overlaps do not seem to have as many names, but the alternative terms for overlaps or durations of overlaps include, at least, *double talking* and *(negative) response times* (Norwine & Murphy, 1938), *double talk* and *interruptions* (Brady, 1968), *simultaneous speech* (Jaffe & Feldstein, 1970), *(negative) switch time* or *switch overlaps* (Sellen, 1995), and *(negative) floor transfer offsets* (de Ruiter et al., 2006). Apparently, there are two ways of treating gaps and overlaps in the previous literature. Either gaps and overlaps are treated as entirely different "creatures", or they are conceptualized as two sides of a single continuous metric (with negative values for overlaps, and positive values for gaps) that measures the relationship between one person ending a stretch of speech and another starting one (de Ruiter et al., 2006; Norwine & Murphy, 1938; Sellen, 1995). Regarding the terminology for pauses (in the sense of silences or durations of silences within the speech of one speaker) finally, these have also been called *resumption times* (Norwine & Murphy, 1938) and the slightly expanded version *within-speaker pauses*.

On a side note, while many of these terms superficially appear to presuppose the existence of turns or a conversational 'floor', studies involving larger scale distribution analyses of such durations have typically defined their terms operationally in terms of stretches of speech ending in a speaker change, rather than stretches of speech ending in a transition-relevance place (cf. ten Bosch et al., 2005).

In this paper, we will adhere to the terminology of Sacks et al. (1974) for referring to acoustic silences and overlaps in conversations with the minor modifications that we will use *gap* for silences bounded by speech from different speakers rather than taking TRPs into account, and that we will not distinguish lapses from gaps. The term *pause* will be used for acoustic silences bounded by speech by the same speaker. *Overlap* will be used for portions of speech delivered simultaneously with speech from another speaker in a speaker change. In addition, we will use *between-speaker intervals* as a cover term for gaps and overlaps similar to the above mentioned *response times, floor transfer offsets* and *switching times* (de Ruiter et al., 2006; Norwine & Murphy, 1938; Sellen, 1995). Similarly, we will use *within-speaker intervals* as a cover term for pauses and what we will refer to as within-speaker overlaps.

2.2. No gaps and no overlaps

In their seminal paper proposing a model for turn-taking organization, Sacks et al. (1974) devoted considerable attention to the phenomenon of speaker changes. Theoretically, there are three possible ways of organizing a speaker change: there may be a silence in-between; there may be overlap; or there may be neither silence nor overlap. From substantial auditory analyses of conversational data, Sacks et al. (1974) had observed that the most common case in conversation is *one-party-at-a-time*, and that speaker changes typically occur without any silence in-between and without any overlapping speech—no-gap–no-overlap. The tendencies observed furthermore lead them to hypothesize a force acting to minimize gap and overlap in conversation.

As is evident from the following quote, however, Sacks et al. (1974) recognized that slight departures from one-at-a-time, that is brief periods of overlap *more-than-one-at-a-time* or short

silences *fewer-than-one-at-a-time* were also relatively frequent: "Transitions from (one turn to a next) with no gap and no overlap are common. Together with transitions characterized by slight gap or slight overlap, they make up the vast majority of transitions." (Sacks et al., 1974, p. 700). From later work by the same authors, presumably when they started measuring the silences, it seems that transitions with slight gap are considered the most frequent ones. For example, Jefferson, who termed transitions with slight gap the *'Unmarked Next Position' Onset,* noted that "My impression is that of all the transition-place points, this is the most frequently used. A recipient/next speaker does not start up in 'terminal overlap', nor 'latched' to the very point of possible completion, but permits just a bit of space between the end of a prior utterance and the start of his own" (Jefferson, 1984, p. 8). In another passage, Jefferson described this 'unmarked next position' as: "With this 'unmarked next' positioning one doesn't get a sense of a next utterance being 'pushed up against' or into the prior, nor of its being 'delayed'. It simply occurs next" (Jefferson, 1984, pp. 8–9). Furthermore, Schegloff (2000) who used the term *normal value of the transition space* for the same case, quantifies "just a bit of space" as roughly one syllable, corresponding to a silent interval of about 150–250 ms. Others have made similar observations of transitions where there is no perceptible gap between the cessation of speaking by one person and the commencement of speaking by the next person – the so-called *smooth transitions* – while there might be an acoustic silence (e.g. Beattie & Barnard, 1979; Jaffe & Feldstein, 1970; Kendon, 1967). Walker and Trimboli (1982) furthermore estimated that the threshold for detection of between-speaker silences in conversations lies close to 200 ms.

2.3. Timing in turn-taking theories

Two competing theories stand out regarding how turn-taking is achieved. Both specifically explain how next speakers know when to start talking. On the one side, there is the *projection theory* initially proposed by Sacks and co-workers according to which a next speaker anticipates or projects when the current speaker will finish based on structural and contextual information, and then starts talking at the projected turn-ending (Sacks et al., 1974). On the other side, there is the *reaction* or *signal theory* (e.g. Duncan, 1972; Kendon, 1967; Yngve, 1970) stating that a next speaker starts talking as a direct reaction to a signal that the current speaker is finished, or is about to finish.

The notion of no-gap–no-overlap was a key concept in the initial development of the projection theory. It has also been very influential in subsequent work within that framework. It seems, however, that some of the followers of Sacks and colleagues have interpreted no-gap–no-overlap as literally zero gap and zero overlap. This understanding of no-gap–no-overlap has then been taken as support for a stronger claim: that turn-taking must rely entirely on the ability to project upcoming turn-endings. Typically, the argument goes as follows: as human turn-taking is so precise that next speakers (over and over again) manage to start with no gap and no overlap, prosodic or other acoustic turn-taking signals immediately before the silence cannot be of any relevance, simply because there is no time to react to such signals (e.g. Bockgård, 2007; de Ruiter et al., 2006; Levinson, 1983). Instead, it is claimed, next speakers have to rely on a projection based on syntactical features, although attempts have been made to show the importance of sentence-initial F0 values for predicting utterance length in read speech (Liberman & Pierrehumbert, 1984; Prieto, D'Imperio, Elordieta, Frota, & Vigário, 2006; Prieto, Shih, & Nibert, 1996). Importantly, for projection to result in zero gap and zero overlap, next speakers have to project not only *what* the current speaker will say, but also the exact point in time *when* she or he will finish.

Between-speaker interval distributions provide empirical evidence that can support or challenge the claims of precision timing in turn-taking. A strong tendency towards no-gap–no-overlap predicts a unimodal distribution centered on zero with fewer slight gaps and slight overlaps. If on the other hand a distribution with its peak (or peaks in case it is a bi- or multimodal distribution) offset from zero is observed, this would indicate that either the minimization of gaps and overlaps aiming at zero gap and zero overlap is not a strong force, or that the projection of turn-endings (in time) is not as precise as is commonly claimed, or both. A cumulative distribution of gap and overlap durations, furthermore, gives an indication of the proportion of speaker changes that must be based on something other than a reaction to silence.

2.4. Distribution analyses of pauses, gaps and overlaps

An important consideration when analyzing pause, gap and overlap distributions is the factors known to influence those intervals. For gaps, it has been suggested that increased stress (induced in an interview situation designed to elicit information of an intimate and embarrassing nature) is associated with markedly shorter gaps (Jaffe & Feldstein, 1970). Similarly, competitive conversations, for example conversations involving arguments, have significantly shorter gaps than cooperative conversations, such as friendly chats (Trimboli & Walker, 1984). There have also been reports that gap durations tend to increase with cognitive load (see e.g. Cappella (1979), and references mentioned therein). Within the Map Task domain, it has been shown that more complex tasks, lack of familiarity with tasks, and presence of conversational game boundaries results in longer gaps (Bull & Aylett, 1998). Several studies have furthermore observed longer gaps in dialogues where the participants have eye contact than in dialogues without eye contact (Beattie & Barnard, 1979; Bull & Aylett, 1998; Jaffe & Feldstein, 1970; ten Bosch, Oostdijk, & de Ruiter, 2004a). The opposite result, faster speaker changes to the extent that average switching times are negative (or overlapping) have also been observed in eye contact vs. no eye contact comparisons (Sellen, 1995). From analyses of pairs of speakers it has also been suggested that speakers adapt the duration of gaps to those of the other participants, that is a form of accommodation or interlocutor similarity with respect to gap duration (e.g. Edlund, Heldner, & Hirschberg, 2009; Jaffe & Feldstein, 1970; Kousidis & Dorran, 2009; ten Bosch et al., 2005). Finally, it has been observed that the language differences with respect to gap durations seem to be minor (cf. Weilhammer & Rabold, 2003).

A statistical consideration when dealing with gap and overlap durations separately is that distributions of gap durations are typically positively skewed (in part as there is an absolute lower limit, but no real upper limit), and similarly that overlap distributions tend to be negatively skewed. Thus, while arithmetic means may seem close at hand for describing such distributions, they may not present a fair estimation of central tendency. Typically, arithmetical means tend to overestimate the central tendency for gaps as well as for overlaps. It has been observed that gap durations are exponentially distributed (Jaffe & Feldstein, 1970), and in a similar vein that logarithmically transformed gap durations better approximate a Gaussian distribution (e.g. Campione & Véronis, 2002; ten Bosch et al., 2004a; Weilhammer & Rabold, 2003). Hence, other measures of central tendency, including mean durations in the log domain (i.e. geometric means) and medians may be better suited to describe gap and overlap distributions.

Table 1 shows a compilation of gap durations for different languages reported in the literature. Given that some studies reported means (or medians) for untransformed data, whereas others calculated them from log-transformed durations, the figures are not entirely comparable. Whenever possible, the reported values have been complemented with other estimates of central tendency.

A number of studies have also reported data on pause durations, that is, acoustic silences within the speech of one speaker (e.g. Brady, 1968; Jaffe & Feldstein, 1970; Norwine & Murphy, 1938; ten Bosch et al., 2005). Table 2 presents a compilation of such pause duration data. ten Bosch et al. (2005) furthermore distinguished between within-utterance and between-utterance (or continuation) pauses within turns, based on a manual segmentation into utterances and turns.

Finally, distributions of overlaps at speaker changes have been presented in (e.g. Brady, 1968; Jaffe & Feldstein, 1970; Weilhammer & Rabold, 2003). Table 3 presents a compilation of overlap data. Note that the values from studies by Norwine and Murphy (1938), Sellen (1995), and de Ruiter et al. (2006) presented in Table 1 included positive as well as negative values, that is gaps as well as overlaps, and that these results are not replicated in Table 3.

2.5. Minimal response times for spoken utterances

In this study, we relate distributions of between-speaker intervals to minimal response times for spoken utterances. It is generally assumed that almost any human behavior involves processes linking perception, decision-making, and action. This in turn has led to the idea that the time from stimulus to response – the reaction time – can be analyzed as the sum of processing times required for the different stages (see e.g. Posner (2005), and references mentioned therein). Reaction time has been measured for many kinds of human behavior, including the time required

Table 1
Different measures of central tendency for gap durations (in ms) reported in the literature.

Language	Eye cont.	Mean	Median	Geom. mean	Std. dev.	Source
English	No	410	320			Norwine and Murphy (1938)[a]
English	No	345	264		104	Brady (1968)[b]
English	No	507	400			Beattie and Barnard (1979)
English	No	474	333			Beattie and Barnard (1979)
English	No	480			620	Sellen (1995)[c]
English	Yes	664			165	Jaffe and Feldstein (1970)
English	Yes	575	360			Beattie and Barnard (1979)
English	Yes	−460			660	Sellen (1995)[c]
English	Yes			380		Weilhammer and Rabold (2003)
English	?	404			421	Bull (1996)
English	?	384	355			Wilson and Wilson (2005)[d]
French	Yes	629	451	496		Campione and Véronis (2002)
Dutch	No	−78			798	de Ruiter et al. (2006)[e]
Dutch	No	380	330		310	ten Bosch et al. (2005)
German	Yes			363		Weilhammer and Rabold (2003)
Japanese	Yes			389		Weilhammer and Rabold (2003)

[a] The *response times* included positive as well as negative values, i.e. gaps as well as overlaps.
[b] Values obtained using the most sensitive speech detector (threshold at −45 dBm) in their study, less sensitive detectors gave higher values.
[c] The *switch times* included positive as well as negative values, i.e. gaps as well as overlaps.
[d] Mean and median values within ±45 ms estimated from tabular frequency distribution.
[e] The *floor transfer offset* (FTO) values included positive as well as negative values, i.e. gaps as well as overlaps.

Table 2
Different measures of central tendency for pause durations (in ms) reported in the literature.

Language	Eye cont.	Mean	Median	Std. dev.	Source
English	No	730	600		Norwine and Murphy, (1938)
English	No	488		93	Brady, (1968)[a]
English	Yes	596		93	Jaffe and Feldstein (1970)
Dutch	No	300	280	210	ten Bosch, et al. (2005)[b]
Dutch	No	520	450	380	ten Bosch, et al. (2005), ten Bosch, et al. (2004b)[c]

[a] Values obtained using the most sensitive speech detector (threshold at −45 dBm) in their study, less sensitive detectors gave higher values.
[b] Pauses within utterances.
[c] Pauses between utterances within the speech of one speaker.

Table 3
Different measures of central tendency for overlap durations (in ms) reported in the literature.

Language	Eye cont.	Mean	Median	Geom. mean	Std. dev.	Source
English	Yes	413			55	Jaffe and Feldstein (1970)
English	Yes			257		Weilhammer and Rabold (2003)
English	No	280	199		61	Brady (1968)[a]
German	Yes			331		Weilhammer and Rabold (2003)
Japanese	Yes			155		Weilhammer and Rabold (2003)

[a] Values obtained using the most sensitive speech detector (threshold at −45 dBm) in their study, less sensitive detectors gave higher values.

for initiating a vocal response to various stimuli. It seems that the fastest a human can react to some stimulus with a vocal response under maximally favorable conditions is about 200 ms (Fry, 1975; Izdebski & Shipp, 1978; Shipp et al., 1984). These minimal response times were observed in a so-called *simple reaction time paradigm* in which the subjects are instructed to react as quickly as possible to a stimulus by performing a predetermined task—in this case to produce a neutral vowel; no decision-making or discrimination of the stimulus is involved; there is only one stimulus and one response; and reaction time is measured from stimulus onset to response onset.

Longer reaction times can be expected in situations that are more complex than the simple reaction time paradigm. For example, the mean reaction time was 496 ms in an experiment that required discrimination of two stimuli (500 and 2000 Hz tones); that the subjects counted the target stimuli (the 2000 Hz tones); and that they responded with an /a/ as quickly as possible on each tenth occurrence of the target (Ferrand, Blood, & Gilbert, 1991).

Another situation that must be considered more complex than the simple reaction time paradigm is that of 'saying something at an appropriate time'. Wesseling and van Son (2005) proposed that such a task could be analyzed within a reaction time framework and devised an experiment where minimal responses (or back-channels) were used as a means to measure response times from *transition relevance places* (TRPs) (Sacks et al., 1974) under the implicit assumption that minimal responses can only occur at TRPs. Subjects were given the task to act as if they participated in a recorded natural conversation by responding with the minimal response "ah" as often as they could. Thus, although there was only one pre-planned response, there was more than one stimulus (in fact probably as many as there were trials); and discrimination and decision-making based on stimuli were involved as the subjects had to decide if it was suitable to say something or not. The interval from the end of the preceding utterance to the onset of the minimal response was measured, and included negative values (i.e. responses overlapping the end of the preceding utterance) as well as positive ones. Values ranging from -1 to $+1$ s from the onset of the minimal response were included in the analyses. The arithmetic mean of the measured intervals was 102 ms with a standard deviation of 454 ms; the mode of the distribution function as estimated from a published histogram was 125 ± 25 ms (Wesseling & van Son, 2005).

The interval from end of utterance to the onset of the minimal response ("ah") was not taken to be the total processing time needed for 'saying something at an appropriate time', however. Instead, the total processing time was estimated from the measured intervals using a mathematical model for analyzing response times proposed by Sigman and Dehaene (2005). The model is valid for tasks that can usefully be divided into three successive stages: a perceptual stage, a central integration or decision stage based on noisy integration of evidence, and a motor stage. In this model, it is assumed that the perceptual and motor stages can operate in parallel with stages of another task, while the central decision process constitutes a bottleneck. These properties are captured in a random-walk (or drunkard's walk) model where the perceptual and motor stages add fixed delays, while the central stage adds a stochastic delay that is the result of a random walk to a decision threshold with a fixed drift rate and added Gaussian noise (cf. Sigman & Dehaene, 2005). The estimates obtained using the Sigman and Dehaene model indicated that the total processing time needed for 'saying something at an appropriate time' was roughly 400 ms, which in their data meant that the planning on average must have started more than 300 ms before the end of the utterance (cf. Wesseling & van Son, 2005). This would imply that the imminent TRP must either have been signaled by the speaker, anticipated by the listener, or a combination of the two, at least 400 ms before the next speaker begins. Any signals indicating that the speaker is not yet finished despite a disruption of the flow of speech may conceivably occur later, as they are not meant to trigger a response, but rather to inhibit one.

By relating distributions of between-speaker intervals to minimal response times for spoken utterances, we can quantify the proportion of speaker changes where the gap is long enough for the next speaker to react to the offset of speech, to silence or to some prosodic information immediately before the silence.

3. Methods

In the following sections, we will first describe the speech material taken from different languages and datasets; next we will give operational definitions of pauses, gaps and overlaps relying on information that can be extracted automatically and with reasonable reliability from the kind of material we use, and outline how this is done within a computational model of interaction. Finally, we will describe how the pause, gap and overlap durations were extracted and analyzed statistically.

3.1. Materials

We used speech material representing three different languages – Dutch, Swedish and Scottish English – taken from three different corpora. The analyses of Dutch were based on spontaneous telephone and face-to-face conversations from the Spoken Dutch Corpus (e.g. Boves & Oostdijk, 2003). The telephone conversations were mainly friendly chats while the face-to-face recordings also included conversations about games played, or other tasks performed during the recordings (cf. ten Bosch et al., 2005). The Dutch dataset was kindly provided to us by Rob van Son and contained raw data on durations of gaps and overlaps derived from manually verified word segmentations; there was no data on pauses. We have not analyzed the sound files ourselves for the Dutch dataset. The between-speaker interval, as defined in the Dutch data, ranges from -2 s overlap to $+2$ s gap. That is, the next speaker should start within ± 2 s of the offset of the previous talkspurt. The next speaker should also start at least 100 ms after the onset of the previous talkspurt. In total 321 speakers, 177 female and 144 males are represented in the Dutch data. The speakers formed 234 pairs, of which 132 were from face-to-face dialogues and 102 from telephone conversations. While the dataset allows for analyses of differences between, for example, eye contact vs. no eye-contact conditions or gender differences, we chose not to subdivide the dataset to make such comparisons. However, see (ten Bosch et al., 2005, 2004a, 2004b) for a number of such analyses on data drawn from the same source.

The analyses of pauses, gaps and overlaps in Scottish English were based on the original Map task corpus, 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*, has a route marked on his or her map. The other speaker, the *instruction follower*, 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.). The HCRC Map Task Corpus includes recordings of 128 map task dialogues (approximately 15 hours of dialogue), predominantly in Standard Scottish English, the variety of Northern English spoken in Glasgow and in the surrounding area. The dialogues were

recorded under four different conditions: familiar and unfamiliar speakers with and without eye contact. Sixty-four speakers (32 female and 32 male) are represented in the data. Each speaker participated in four dialogues, twice as instruction giver, twice as instruction follower, once in each case with his or her familiar partner, once with an unfamiliar partner. Half of the dialogues were recorded with eye contact and the other half without eye contact (Anderson et al., 1991). There was a good acoustic separation of the speaker channels. As in the case of the Spoken Dutch corpus, the HCRC Map Task Corpus allows for comparisons between different conditions, but also here we chose not to subdivide the dataset. However, see (Bull & Aylett, 1998) for a number of such analyses on these data. Although there are various kinds of mark-up and segmentations available for this corpus, the data on between- and within-speaker intervals presented in the present study were derived by us using a computational model of interaction.

The analyses of pauses, gaps and overlaps in Swedish, finally, were based on the Swedish Map Task Corpus (Helgason, 2002, 2006) designed as a Swedish counterpart to the HCRC Map Task Corpus. Eight speakers, five females and three males, are represented in this 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 h and 18 min. The dialogues were recorded in an anechoic room, using close-talking microphones, with the subjects facing away from each other (i.e. without eye contact), and with acceptable acoustic separation of the speaker channels.

3.2. Procedures

Pauses, gaps and overlaps were operationally defined in terms of a computational model of interaction. This interaction model is computationally simple yet powerful and uses boundaries in the conversation flow, defined by the relative timing of speech from the participants in the conversation, as the only source of information. In particular, 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, 1987; Jaffe & Feldstein, 1970; Laskowski & Shriberg, 2009; Norwine & Murphy, 1938; Raux & Eskenazi, 2009; 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. This distinguishes our model from the ones where manual annotations are used to identify turns continuing across silences and intervening speech from other speakers, and where these labels are subsequently used to classify these cases as within-speaker events (e.g. Sellen, 1995; ten Bosch et al., 2005; Weilhammer & Rabold, 2003). We use the resulting conversation state labels to extract the durations of the states. The procedure involves three steps, as depicted in Fig. 1.

First, we perform vocal activity detection, individually for each speaker, using VADER from the CMU Sphinx Project ("The CMU Sphinx Group Open Source Speech Recognition Engines," n.d.). This produces a labeling of each 10 ms frame, for each speaker, as either SPEECH or SILENCE, resulting in a maximum temporal resolution of 10 ms. VADER bridges silences of less than 180 ms, so that the smallest pause duration present in VADER output is 180 ms. Gap durations, on the other hand, are defined as silences between the offset of one person's speech and the onset of another's, making the minimal detectable gap duration 10 ms, see Fig. 2. The 180 ms bridging minimizes the risk of mistaking stop closures for pauses. An analysis of automatic segmentations of about 13,000 voiceless stops collected within the GROG project (see e.g. Heldner & Megyesi, 2003; Sjölander & Heldner, 2004) showed that 99.2% of the stop closures had a duration of less than 180 ms. VADER also bridges very short stretches of speech, so that any detected talkspurt of less than 90 ms is removed which minimizes the risk for mistaking noise (e.g. raps, knocks) for speech.

Second, at each instant, the SPEECH and SILENCE 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 speaker 1 (SP_1), the state is SELF if SP_1 is speaking and speaker 2 (SP_2) is not; it is OTHER if SP_1 is silent and SP_2 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 2 is similar; we illustrate this process for both speakers in the middle panel of Fig. 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

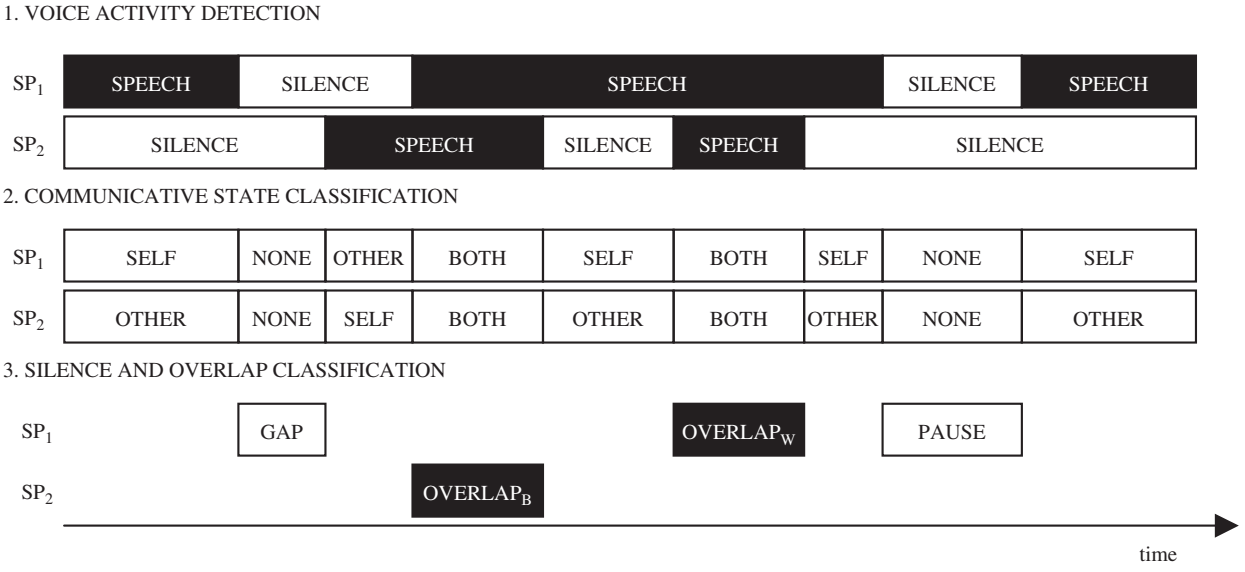

Fig. 1. Illustration of how gaps, overlaps ($OVERLAP_B$), pauses, and within-speaker overlaps ($OVERLAP_W$) are defined and classified in the interaction model. The illustration shows all three steps (as in the text) from the perspectives of both speaker 1 (SP_1) and speaker 2 (SP_2).

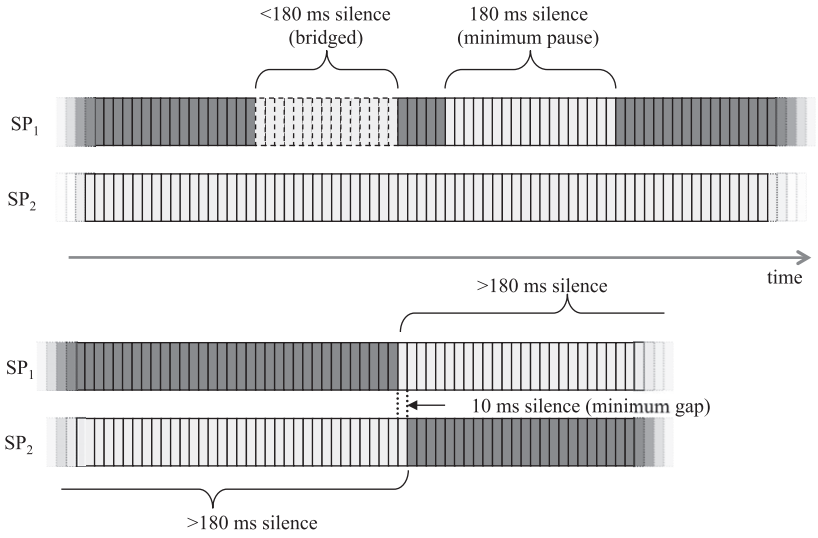

Fig. 2. Illustration of bridging of silence, minimum pause, and minimum gap in the interaction model. A grey frame represents 10 ms of detected speech; a white frame represents 10 ms of detected silence. SP1 and SP2 represent two speaker channels.

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 (due to the temporal resolution this means in effect with a silence or overlap smaller than 10 ms) 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 speaker 1 or speaker 2 are speaking alone, can be modeled as 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.

These operational definitions represent only minor modifications with respect to the original definitions by Sacks et al. (1974). A BETWEEN-SPEAKER SILENCE in this model corresponds to a *gap*; a BETWEEN-SPEAKER OVERLAP corresponds to an *overlap*; a WITHIN-SPEAKER SILENCE corresponds to a *pause*; whereas a within-speaker overlap has no direct correspondence in the terminology of Sacks et al. (1974) as far as we understand. We did not analyze the within-speaker overlaps in the present study.

Once the pauses, gaps and overlaps were identified and classified, their durations were extracted by subtracting the time of the onset of an interval from the time of its offset. Subsequently, several types of explorative statistical analyses including histograms, cumulative distributions, percentile ranks and descriptive statistics were performed using the statistical analysis software SPSS.

4. Results

4.1. Data splits and transformations

Two important considerations when calculating or comparing between- and within-speaker intervals from different studies are (i) whether the between-speaker intervals are split into gaps and overlaps or not, and (ii) whether the durations are logarithmically transformed or not. We will exemplify this point with analyses of about 43,000 between-speaker intervals from the Spoken Dutch Corpus (e.g. Boves & Oostdijk, 2003; ten Bosch et al., 2005). Fig. 3 shows histograms of between-speaker intervals split into overlaps (intervals < -10 ms) and gaps (intervals > 10 ms). Table 4 shows corresponding descriptive statistics for the split categories as well as for the combined data. Table 4 also contains 169 no-gap–no-overlap cases (i.e. -10 ms $<$ intervals $<$ 10 ms) that are not shown in Fig. 3.

First, these analyses showed that the distributions of overlaps, gaps, as well as the distribution of all between-speaker intervals deviated substantially from symmetry and normality as indicated by skewness and kurtosis values of more than twice their respective standard errors. Hence, arithmetic means do not provide particularly meaningful estimates of central tendency for these distributions. Compare for example the peak of the gap distribution with the peak of the overlaid normal distribution (i.e. the mean) in the right panel of Fig. 3. Medians or geometric means appear to give more realistic estimates of central tendency. Furthermore, it makes a considerable difference whether the estimates of central tendency are calculated from all between-speaker intervals (as in de Ruiter et al., 2006; Norwine & Murphy, 1938; Sellen, 1995) or from gaps and overlaps separately (as in e.g. Brady, 1968; Jaffe & Feldstein, 1970; Weilhammer & Rabold, 2003). Obviously, smaller estimates of central tendency are to be expected when overlaps are included in the distributions compared to gap only distributions (cf. Table 4).

Logarithmic transformation of duration data has often been suggested as a means of making duration data less skewed and better described by a normal distribution (e.g. Campione & Véronis, 2002; Jaffe & Feldstein, 1970; ten Bosch et al., 2004a; Weilhammer & Rabold, 2003). Fig. 4 shows histograms of overlap and gap durations for the same data as in Fig. 3 plotted on a logarithmic (base 10) scale.

Visually, it appears that both the overlap and the gap distributions are better described by a normal distribution when plotted on a logarithmic scale. Inspection of skewness and kurtosis values calculated from logarithmically transformed absolute durations of gaps and overlaps, however, do not warrant such a conclusion. The overlap distribution even got more skewed (skewness -1.2 vs. -0.9) and more leptokurtic (kurtosis 1.8 vs. -0.1) than it was before. The gap distribution got slightly less skewed, although the polarity of the skewness changed (-1.0 vs. 1.6).

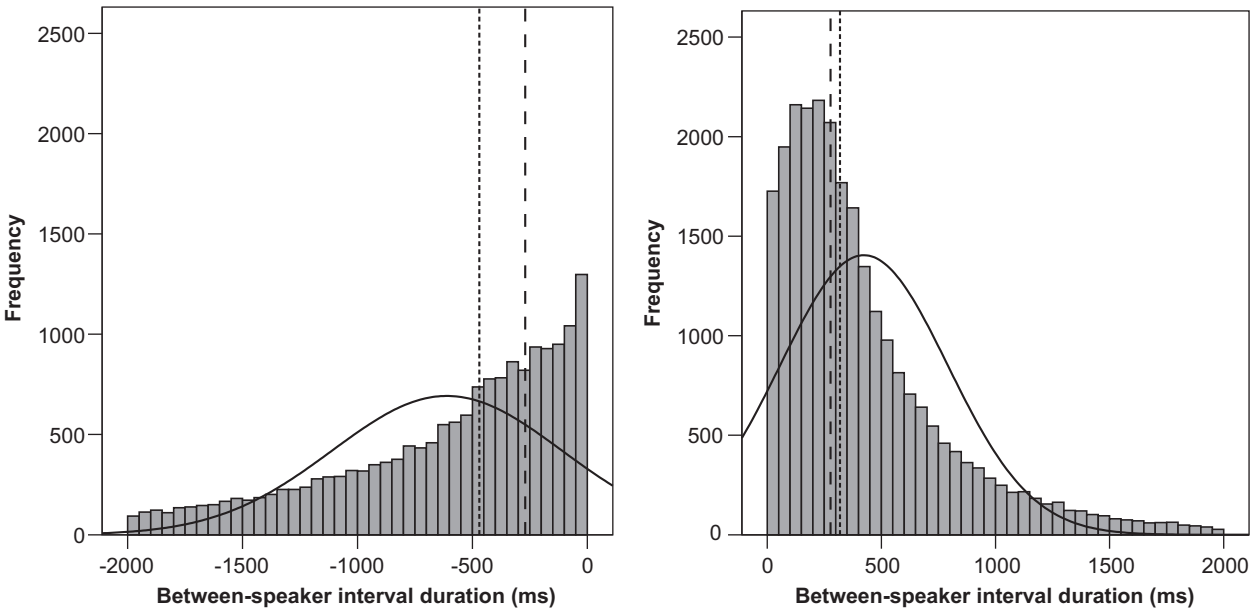

Fig. 3. Histograms of between-speaker interval durations (in ms) split into overlaps (left) and gaps (right) from the Spoken Dutch Corpus with overlaid normal distributions. Bin size 50 ms. Dashed lines indicate the medians, and stitched lines the geometric means (calculated from absolute durations) for the respective distributions.

Table 4
Descriptive statistics for the durations of overlaps, no-gap–no-overlaps, gaps, as well as for all between-speaker intervals (in ms). Data from the Spoken Dutch Corpus.

	Overlaps	No-gap–no-overlaps	Gaps	All
Mean	−610	0	424	8
Median	−470	0	318	111
Geometric mean	−372[a]	0	277	
Skewness	−0.9		1.6	−0.5
Std. error of skewness	0.02		0.02	0.01
Kurtosis	−0.1		2.5	0.7
Std. error of kurtosis	0.04		0.03	0.02
N	17,361	169	25,844	43,374
% Of total *N*	40.0	0.4	59.6	100.0
Minimum	−2000	0	1	−2000
Maximum	−1	0	1999	1999

[a] Geometric mean calculated from the absolute values of overlap but expressed as a negative value to conform to the rest of the numbers in the overlap column.

What was more alarming with this log transform exercise, however, was that whereas the distribution of untransformed data appeared unimodal with one clear peak at approximately 200 ms gap (cf. Fig. 3), the same data plotted on a logarithmic scale gave the impression of a bimodal distribution with peaks at approximately −400 ms (overlaps) and 300 ms (gaps), see Fig. 4. This bimodality was clearly an artifact of the transformation, and can to a large extent be explained by lower bin counts as a result of the narrowing bin widths (in untransformed durations) as the log-transformed duration approached zero.

As the Spoken Dutch Corpus represents a large dataset, we feel that these observations may be relevant also for other datasets. In the following analyses, we will treat gap and overlap durations as one distribution—as a distribution of between-speaker intervals, and will not transform the durations. As a general recommendation, we suggest that whenever gap as well as overlap durations are available, they should be treated as one distribution, and that no transformation should be applied. Transformation might be relevant in the case that only one side of the between-speaker interval distribution is available and you want to make analyses that depend on a normal distribution. This may be the case for example in a reactive system such as a silence based end-of-utterance detector used in a spoken dialogue system. As observed above, however, normally distributed data cannot be guaranteed from a log transform.

4.2. Between-speaker intervals

Table 5 presents frequencies and percentages of the different types of between-speaker intervals in the Spoken Dutch Corpus, the HRCR Map Task Corpus, and the Swedish Map Task Corpus, respectively. Figs. 5–7 presents histograms and cumulative distributions of between-speaker interval durations (gaps and overlaps as one distribution) in the three corpora. Table 6 presents descriptive statistics for between-speaker intervals in the same corpora.

These analyses showed that no-gap–no-overlap in the strict sense of between-speaker intervals ranging from −10 to 10 ms clearly was not the most frequent type of between-speaker interval, but rather a very rare one. These no-gap–no-overlaps represented less than 1% of the between-speaker intervals in our data. The most frequent kind of between-speaker interval was instead a *slight gap.* The mode of the distribution function was offset from zero by about 200 ms in the histograms for all three corpora, and the different measures of central tendency for the distributions were all on the gap side. These analyses also showed that all three distributions deviated substantially from a normal distribution, both in terms of skewness and of kurtosis (i.e. skewness and kurtosis values of more than twice their respective standard errors), which partly explains the discrepancies between means and medians. Furthermore, the deviations from no-gap–no-overlap to the negative end of the scale, that is overlaps, were also frequent. The overlaps represented about 40% of all between-speaker intervals in our material.

When we instead looked at the cumulative distribution below the threshold for detection of between-speaker intervals, that is the distribution of *smooth transitions* or transitions without

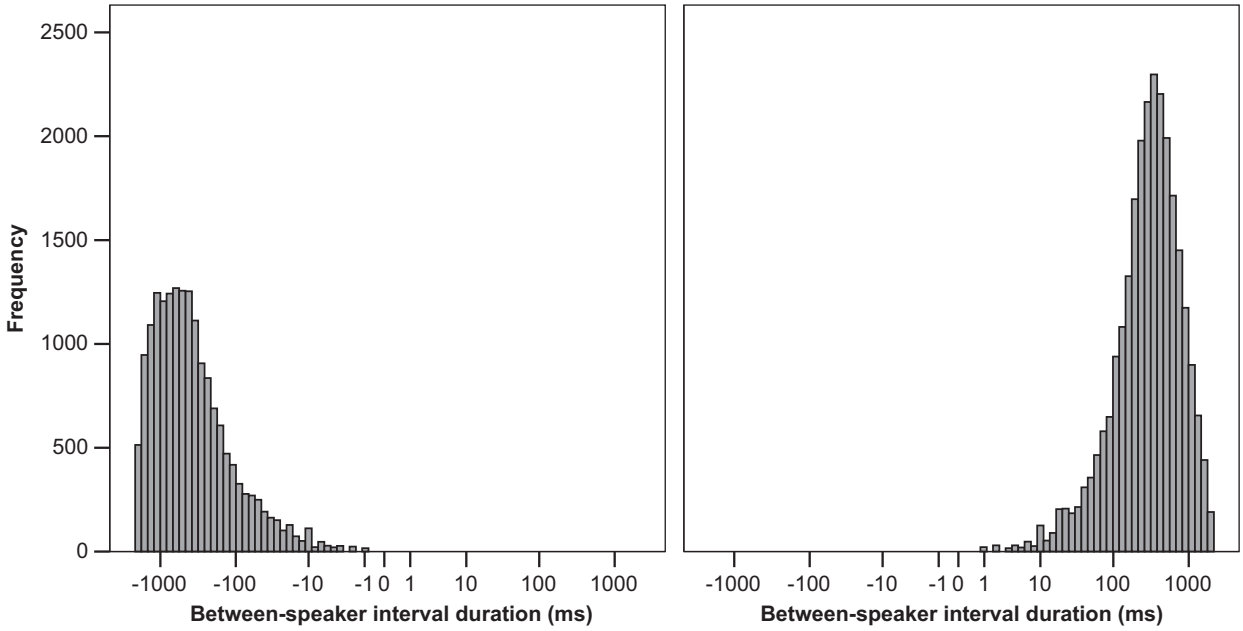

Fig. 4. Histograms of overlap (left) and gap (right) durations (in ms) plotted on a logarithmic (base 10) scale. Data from the Spoken Dutch Corpus.

Table 5
Frequencies and percentages of the different types of between-speaker intervals in the Spoken Dutch Corpus (CGN), the HCRC Map Task Corpus (MTC), and the Swedish Map Task Corpus (SMTC).

	Frequency			Percent		
	Gap	Overlap	No-gap–no-overlap	Gap	Overlap	No-gap–no-overlap
CGN	25,844	17,361	169	59.6	40.0	0.4
MTC	8915	6457	115	57.6	41.7	0.7
SMTC	1225	824	11	59.5	40.0	0.5

perceptible gap (e.g. Walker & Trimboli, 1982), we observed that about 14–19% of all gaps were shorter than 200 ms, and furthermore that 55–59% of all between-speaker intervals were either not noticeable gaps, or overlaps. Unfortunately, we are not aware of any studies of any corresponding threshold of detection for overlaps.

The cumulative distribution above the 200 ms threshold was also of interest, as it represented the cases where reaction to cessation of speech might be relevant given published minimal reaction times for spoken utterances (Fry, 1975; Izdebski & Shipp, 1978; Shipp et al., 1984). The distribution above this threshold represented 41–45% of all between-speaker intervals. These cases were thus potentially long enough to be reactions to the cessation of speech, or even more so to some prosodic information just before the silence.

We also observed that 70–82% of all between-speaker intervals (i.e. gaps and overlaps) were shorter than 500 ms, and similarly that 82–95% of all intervals were shorter than 1000 ms. Acoustic silence thresholds at 500 ms or 1000 ms are used in many end-of-utterance detectors in speech technology applications. Consequently, a speech technology application using a 500 ms acoustic silence threshold would have captured only 18–30%, and a 1000 ms threshold only 5–18% of all between-speaker intervals in the kind of conversations represented in our material.

Generally, the similarities between the three datasets with respect to the proportion of gaps and overlaps, the location of distribution function modes as well as the general shape of the distributions were striking, despite the fact that they represented different languages as well as slightly different kinds of conversations.

4.3. Within-speaker intervals

Table 7 presents a comparison of selected descriptive statistics for within- and between-speaker silence durations (i.e. pauses and gaps) in the Swedish Map Task Corpus and in the HCRC Map Task Corpus. The Spoken Dutch Corpus is not included here, as we do not have any pause data from that dataset.

This comparison revealed that, at least the way we define and extract them, there were relatively more pauses than there were gaps in both these corpora. Furthermore, pauses generally had longer durations than gaps, no matter what measure of central tendency was contrasted between the two.

Furthermore, an examination of the proportion of pauses and gaps with durations of more than 500 ms, a common silence threshold in end-of-utterance detectors, showed that such a threshold captured 51.1% and 47.5% of all gaps, but also 59.6% and 56.0% of all pauses in the Swedish Map Task Corpus and the HCRC Map Task Corpus, respectively. With a 1000 ms silence threshold, the corresponding values were 29.1% and 25.1% for gaps, and 31.0% and 27.7% for pauses, for the two corpora. As there were more pauses than gaps, both silence duration thresholds captured more pauses than gaps also in absolute numbers.

5. Discussion

5.1. Turn-taking is less precise than is often claimed

This study indicates that the timing of turn-taking is not as precise as is often claimed. Speaker changes are not strictly no-gap–no-overlap and one-speaker-at-a-time. Instead, sizeable departures from no-gap–no-overlap occur frequently, while cases with neither gap nor overlap are very rare. The most common between-speaker interval in all three examined corpora, as

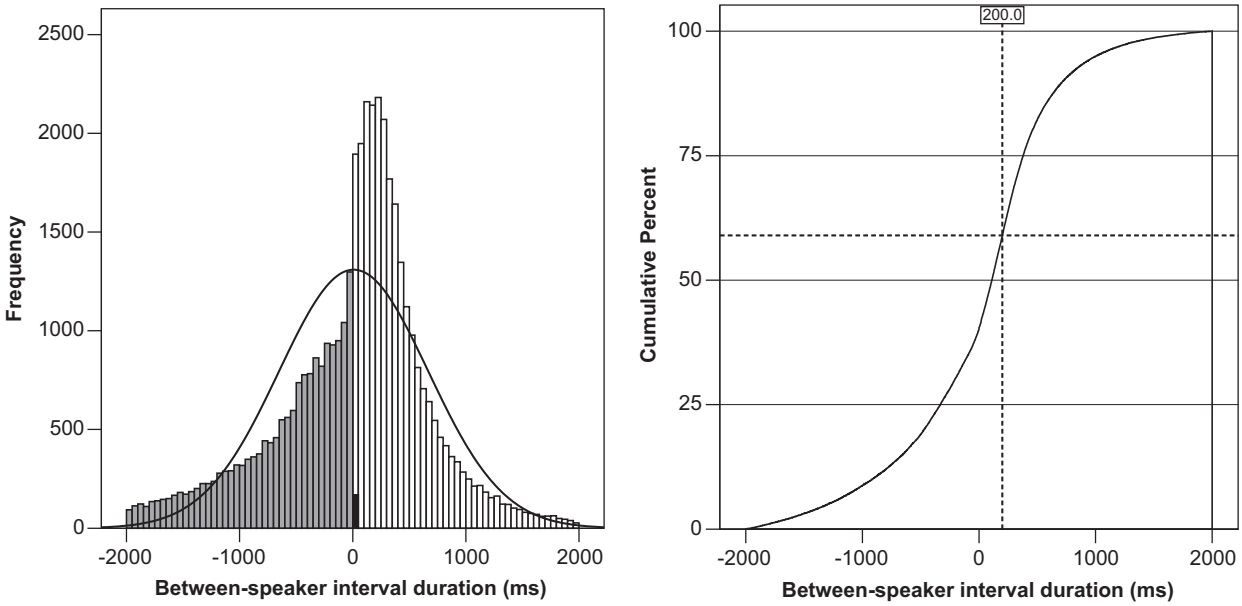

Fig. 5. Histogram and cumulative distribution of between-speaker intervals (in ms) from the Spoken Dutch Corpus. Bin size 50 ms. The vertical dashed line in the right panel shows the threshold for detection of gaps (200 ms), and the horizontal one the cumulative distribution up to that threshold (59%).

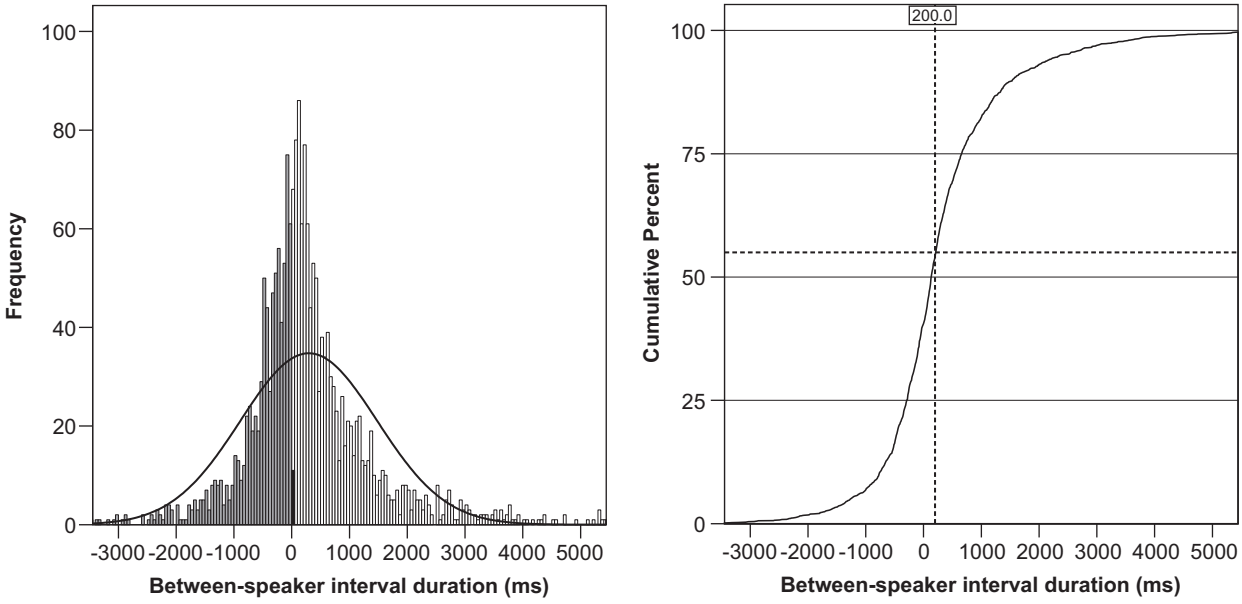

Fig. 6. Histogram and cumulative distribution of between-speaker intervals (in ms) from the HCRC Map Task Corpus. Bin size 50 ms. The vertical dashed line in the right panel shows the threshold for detection of gaps (200 ms), and the horizontal one the cumulative distribution up to that threshold (57%).

indicated by the modes of the distribution functions, is a gap of about 200 ms. That is, the most frequent between-speaker interval is a *slight gap*, or a *just noticeable gap* (e.g. Jaffe & Feldstein, 1970; Walker & Trimboli, 1982). Gaps with a duration above the threshold for detection of silences represent more than 40% of all between-speaker intervals in our material. Measures of central tendency in this range, or slightly above, is also what has been observed in most previous studies (cf. e.g. Jefferson, 1984; Schegloff, 2000), with a few exceptions (e.g. de Ruiter et al., 2006; Sellen, 1995).

Overlaps, which constitute the other possible deviation from no-gaps-no-overlap and one-speaker-at-a-time, are also frequent. The overlaps represent about 40% of all between-speaker intervals in our material. Unfortunately, we are not aware of any studies that have determined the duration threshold for detection of overlaps, which is why we just report the proportion of intervals below −10 ms. The proportion of overlaps may seem high given previous accounts citing less than 5% overlap (Levinson, 1983), or about 8% overlap (Norwine & Murphy, 1938). On the other hand, it is lower than the studies reporting, for example, 44% overlap in face-to-face dialogues and 52% in telephone dialogues (ten Bosch et al., 2005), or 54.1% overlap in a four-party conversation with people sitting around a table (Sellen, 1995).

From these observations, we argue that interlocutors do not set their aim strictly at one-speaker-at-a-time and no-gap–no-overlap. These cases represent only a marginal part of our data. Turn-taking is a highly practiced skill, and it simply cannot be the case that the vast majority of attempts to take the turn in effect

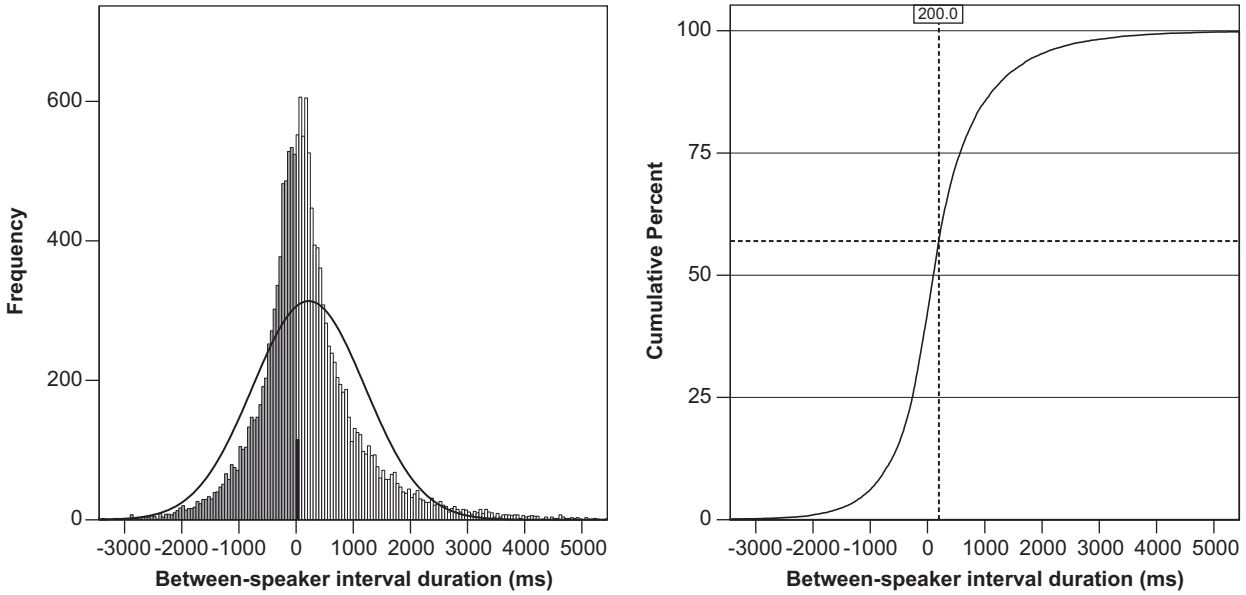

Fig. 7. Histogram and cumulative distribution of between-speaker intervals (in ms) from the Swedish Map Task Corpus. Bin size 50 ms. The vertical dashed line in the right panel shows the threshold for detection of gaps (200 ms), and the horizontal one the cumulative distribution up to that threshold (55%).

Table 6
Descriptive statistics for between-speaker intervals (in ms) in the Spoken Dutch Corpus (CGN), the HCRC Map Task Corpus (MTC), and the Swedish Map Task Corpus (SMTC).

	CGN	MTC	SMTC
Mean	8	223	295
Std. deviation	661	985	1182
Median	111	110	130
Skewness	−0.5	1.4	2.3
Std. error of skewness	0.01	0.02	0.05
Kurtosis	0.7	8.2	18.4
Std. error of kurtosis	0.02	0.04	0.11
N	43,374	15,487	2060

Table 7
Comparison of descriptive statistics of between- and within-speaker silence durations (in ms) in the Swedish Map Task Corpus (SMTC) and in the HCRC Map Task Corpus (MTC).

	SMTC		MTC	
	Gap	Pause	Gap	Pause
Mean	888	990	766	872
Median	520	640	470	580
Geometric mean	458	683	418	632
N	1225	1496	8915	12,002

miss the goal. If we relax no-gap–no-overlap slightly to include not noticeable gaps (i.e. gaps up to 200 ms), the resulting proportion still represents a minority of about 20% of all between-speaker intervals. Consequently, if the hypothesized force acting to minimize gaps and overlaps (Sacks et al., 1974) exists, it is not a very strong one. On a side note, an estimate of the threshold for detection of overlaps similar to the threshold for detection of gaps determined by Walker and Trimboli (1982), would allow us to estimate the proportion of *not noticeable gaps and not noticeable overlaps*, which is perhaps closer to what Sacks et al. (1974) intended for the no-gap–no-overlap principle.

Assuming instead that we, as highly trained speakers, succeed more often than we fail at turn-taking, slight gaps is a more plausible goal for between-speaker intervals. We note, however, that a turn-taking model assuming precision-timed speaker changes aimed at a constant between-speaker interval, be it 0 ms or 200 ms, does not fit the observed data, as the data is highly distributed.

From anecdotic data and introspection, we note that a reasonable gap duration in one situation can be awkward in another. It is perfectly normal to respond to a greeting after only a slight gap, but delaying the response for a second or two will alter its meaning. Conversely, a response to a complex question is going to sound disturbingly insincere if delivered too soon. It is likely that there are similar tendencies for overlaps. For example, overlapping the end of a highly predictable utterance may be entirely acceptable, whereas overlap into completely unpredictable content may be disturbing or rude. Whether the predictability of an utterance and its speech act are key factors remains to be investigated, but it is clear that there are other factors at play than a single drive towards one-speaker-at-a-time.

5.2. Implications for timing in turn-taking theories

Distributions of between-speaker intervals can be used in arguments both for projection theory (e.g. Bockgård, 2007; de Ruiter et al., 2006; Sacks et al., 1974) and for reaction or signal theory (e.g. Duncan, 1972; Kendon, 1967; Yngve, 1970). On the one hand, this study presents evidence speaking against the signal or reaction theory or at least calling for alternative or complementary explanations, such as those offered by the projection theory. This study has quantified the proportion of cases where the next speaker's decision to start speaking cannot possibly be a reaction to interaction control signals near the end of the current speaker's speech. There are indeed cases where it must be assumed that other information than signals such as offset of speech, acoustic silence or intonation patterns immediately before the offset of speech, is required. These cases include speaker changes with overlap, as well as those with gaps shorter than the minimal response time for spoken utterances. We have shown that the cases that cannot rely on reaction to such signals

are frequent in conversations. The proportion of between-speaker intervals below 200 ms in our material ranged from 55% to 59% for the different speech materials. These numbers are higher than those in previous studies reporting that 35% of all between-speaker intervals (including gaps as well as overlaps) were shorter than 200 ms (Norwine & Murphy, 1938), and similarly that 30–34% of all gaps fell below a 200 ms threshold (e.g. Beattie & Barnard, 1979; Brady, 1968; Wilson & Wilson, 2005).

On the other hand, we have also shown that a substantial share of all speaker changes involve gaps long enough for the next speaker to react to potential signals occurring in the immediate vicinity of the speaker change, showing that reaction theory can explain a substantial share of speaker changes. The proportion of between-speaker intervals exceeding the 200 ms gap threshold ranged from 41% to 45% in our material. Thus, acoustic silences (as they become noticeable), the offset of speech (which probably is a more salient perceptual event than the acoustic silence in itself due to its greater spectral change) or intonation patterns just before the silence are all potentially useful as interaction control signals in a little more than 40% of all speaker changes. There may of course also be signals located earlier, such as the proposed *TRP-projecting accents* situated on the last major accented syllable and marking the onset of a TRP interval (Wells & MacFarlane, 1998), or the initial F0 values used in attempts to predict sentence length (Liberman & Pierrehumbert, 1984; Prieto et al., 1996, 2006).

The 200 ms gap threshold used here is based on a minimal response time estimated under maximally favorable conditions. We argue that as interlocutors are highly trained to find suitable places to say something, relating gaps to minimal response times is reasonable. Higher thresholds, such as the 400 ms estimated to be the total processing time for a minimal response (Wesseling & van Son, 2005) would lower the proportion of cases that can be explained by reaction theory, but would clearly not eliminate them. From these observations, we conclude that reaction to interaction control signals is a plausible explanation for a significant proportion of all speaker changes in human–human conversation, and furthermore that a reactive model using acoustic features in the immediate vicinity of the speaker change is a viable alternative for speech technology applications.

Among the proponents of the projection theory, it is often claimed that turn-endings *must* be projectable (i.e. their occurrence in time must be predictable): "for it is this [projection] alone that can account for the recurrent marvels of split-second speaker transition" (Levinson, 1983, p. 297). This study indicates that to the extent that projection is involved, the projections in time in conversation are imprecise – we reiterate that overlaps as well as noticeable gaps occur frequently in our material, while the no-gap–no-overlap cases are rare. Thus, the no-gap–no-overlap principle (Sacks et al., 1974) can neither be used as a part of an argument in favor of projection nor against reaction simply because the no-gap–no-overlap cases hardly ever occur in real speaker change data. Importantly, this means that a principal motivation for projection in turn-taking is invalid. Sacks and co-workers knew this as soon as they started measuring between-speaker intervals (e.g. Jefferson, 1984; Schegloff, 2000). The fact that this argument keeps returning can only be understood as the followers of Sacks et al. (1974) taking the no-gap–no-overlap principle as gospel, and failing to notice details present already in the paper coining this very principle.

A complete model of the control of the interaction, whether it is a model of human behavior or a model for speech technology applications, should of course be able to deal with speaker changes with overlaps and slight gaps as well as with those with noticeable gaps. A reactive model using signals in the immediate vicinity of the speaker change alone will not suffice to achieve this. We argue that these circumstances do not warrant the conclusion that reactive models should be excluded altogether, or that projection models explain the data. While the present study does not give any insights as to whether reaction is in fact used by human interlocutors, we would stress that the possibility of humans using reaction cannot be ruled out based on the observed between-speaker interval data. Furthermore, the fact that reaction cannot explain all cases does not imply that projection explains anything at all, including how next speakers time their speech onset. There may be other explanations, and several mechanisms may operate in parallel.

5.3. A sketch of a synthesis of reaction and prediction

In relation to the reaction vs. projection debate, it is worth stating that we do not consider projection and reaction to turn-ending signals to be mutually exclusive. Redundancy is a well-studied and recurring principle of human language in use on virtually every level, and it is likely that a phenomenon as important as the taking of turns is orchestrated by a number of redundant control methods. We speculate that reaction and prediction are both important in regulating turn-taking.

The idea of projection in turn-taking most likely stems from the everyday observation that we are often able to predict *what* other people are going to say. Predicting the actual point in time *when* a speaker will cease speaking from a prediction of content constitutes an additional and rarely discussed step. The only accounts of human precision timing we are aware of, however, concern phenomena governed by continuous and unvarying forces. Our ability to catch thrown objects, for example, relies on the principles of ballistics involving gravitation, speed and a negligible amount of friction. We are not aware of any studies demonstrating our ability to make predictions of the duration of units of speech (utterances, syllables, etc.) down to fractions of a second. This leads us to conjecture that projection is about content and understanding, rather than about timing.

Projections enable us to formulate responses in advance, so that we do not have to do this after we decide to respond, but they have not been shown to provide precise timing in turn-taking. In fact, we suspect that projection of content may be responsible for a fair share of the speaker changes involving overlap—that is, cases that could be described as less precise with respect to timing of turn-taking. Many overlaps occur because the next speaker is confident about what the current speaker will say, and deliberately responds before the current speaker finishes. Speaker changes often occur when the current utterance becomes predictable in the eyes of the next speaker, so with respect to timing, projection of content may result in overlaps just as well as in gaps. In this interpretation, reaction is used when the continuation of an utterance is not predictable, or, when the next speaker for some reason wishes to wait until the current speaker has finished and stopped talking. Successful reaction, then, can only result in gaps.

We can only speculate about the information required to capture all possible cases, but projection based on signals located earlier relative to the speaker change; direct reaction to signals located earlier with respect to the speaker change; reaction to fulfillment of expectations of semantic content; as well as reaction to projectability of the current utterance (i.e. a point where the rest of the utterance appears predictable to the listener) are good candidates.

5.4. Implications for speech technology

Regarding potential interaction control signals, the present study corroborates the idea of exploring acoustic features in the

immediate vicinity of speaker changes as interaction control signals. There is a possibility that such signals occur, as a substantial proportion of the speaker changes in our material involve gaps long enough for the next speaker to react to some kind of interaction control signals in the immediate vicinity of the speaker change.

Acoustic silences by themselves cannot be considered essential as interaction control signals, however. This study confirms previous findings that the presence or duration of acoustic silences is not particularly informative as to whether a speaker change will occur or not (e.g. Edlund & Heldner, 2005; Ferrer, Shriberg, & Stolcke, 2002). The primary reason for this is the fact that pauses often are longer than gaps, but there is also the fact that speaker changes without silences and with overlaps occur relatively frequently. Furthermore, from the point of view of perceptual relevance, we find it unlikely that acoustic silences are as salient perceptual events as for example the transition from speech to silence (the offset of speech), as the perceptual system is generally better at detecting change than static conditions, and as such a transition involves a large spectral change.

This study also confirms previous observations that the vast majority of between-speaker intervals in conversations are shorter than the 500 or 1000 ms silence thresholds used in many end-of-utterance detectors in speech technology applications. Thus, a silence-only turn-taking decision in a speech technology application will result in significantly fewer speaker changes than if human interlocutors would have made the turn-taking decisions in the same situation. In addition, silence duration thresholds will in many cases result in systems responding slower than humans would have done. The remaining between-speaker intervals will be *longer* than this typical threshold, so there will be occasional cases where a system responds faster than a human would have done in the same situation. While this may seem like a trivial observation, we think it is important to understand that between-speaker intervals are indeed distributed, and that a model of interaction control aiming at human-like behavior must capture also this characteristic. Although systems are generally too slow, they may also respond too fast. Informal tests in a human–computer interaction setting carried out in our lab suggest that very short gaps before system responses can sometimes be perceived as highly disturbing.

Furthermore, we and others have noted that pauses generally tend to have longer durations than gaps (cf. e.g. Brady, 1968; Norwine & Murphy, 1938; ten Bosch et al., 2005). There are also a couple of observations going in the opposite direction: slightly shorter pauses than gaps (Jaffe & Feldstein, 1970); and slightly shorter pauses (within manually labeled utterances) than gaps (ten Bosch et al., 2005). From a speech technology point of view, this implies that reactive models relying solely on silence duration will often cause a system to interrupt its users, namely in situations when the users intended to make a pause rather than a gap (e.g. Edlund & Heldner, 2005; Ferrer et al., 2002). Thus, there are several reasons why silence-only turn-taking decisions are not well suited for speech technology applications, and especially not for systems aiming at more human-like conversational behavior.

This study indicates areas for improvements in reactive end-pointing for speech technology applications. We have shown that about 40% of all between-speaker intervals in genuine conversations are long enough for the next speaker to react to information immediately before the silence given published minimal response times for spoken utterances (Fry, 1975; Izdebski & Shipp, 1978; Shipp et al., 1984). If there is a possibility that humans use reaction in turn-taking, reaction may also be a viable alternative in speech technology. It so happens that current voice activity detection algorithms need to work with a latency (or look-ahead) that is close in duration to the threshold for detection of silences in humans. Thus, speech technology applications could potentially react to the same signals as humans in these cases. For example, when a voice activity detector enters a silence state, the speech immediately before the silence may be inspected for information which informs turn-taking decisions. There is a substantial body of research indicating that prosody is relevant for signaling speaker changes and other interaction control phenomena (see e.g. Edlund & Heldner, 2005, and references therein). Improvements over current technologies in 40% of all speaker changes is interesting enough. Add to that the possibility of avoiding taking the turn in unsuitable silences—pauses. While we are aware that such a model cannot reach human performance, we argue that for speech technology applications, any model outperforming an end-point detector relying on acoustic silence duration only is a step in the right direction. Prosodic interaction control signals, as well as information related to semantic completeness, gaze, gestures, breathing behavior, voice quality, etc. should all be exploited to improve the performance of reactive turn-taking behavior in speech technology whenever reliable estimates of these are available.

6. Conclusions

In this study, we have described and used a zero-manual-effort methodology for the explorative study of durational aspects of turn-taking. Based on analyses of three different conversational corpora representing three different languages, it is shown that the timing of turn-taking is less precise and more distributed than is often claimed. From these observations, we conclude that the target with respect to timing of turn-taking cannot be one-speaker-at-a-time and no-gap–no-overlap, and furthermore that precision timing in turn-taking can neither be used in arguments in favor of projection, nor against reaction as models of timing in turn-taking. Furthermore, as more than 40% of all between-speaker intervals are long enough for the next speaker to react to information immediately before the silence given minimal response times for spoken utterances, we also conclude that reaction is a plausible explanation in a significant proportion of all speaker changes. This in turn, is taken to corroborate the idea of using prosodic features in the immediate vicinity of speaker changes to inform and improve interaction control decisions in spoken dialogue systems and other speech technology applications.

Acknowledgments

We thank Kornel Laskowski for valuable comments, Rob van Son for supplying us with between-speaker interval data from the Spoken Dutch Corpus, Pétur Helgason for access to the Swedish Map Task corpus and Jean Carletta for helpful assistance with the HCRC Map Task Corpus. This work was funded by the Swedish Research Council (VR) Project no. 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,* 83–97.

Beattie, G. W., & Barnard, P. J. (1979). The temporal structure of natural telephone conversations (directory enquiry calls). *Linguistics, 17,* 213–229.

Bockgård, G. (2007). Syntax och prosodi vid turbytesplatser: Till beskrivningen av svenskans turtagning. In E. Engdahl, & A.-M. Londen (Eds.), *Interaktion och kontext: Nio studier av svenska samtal* (pp. 139–183). Lund: Studentlitteratur.

Boves, L., & Oostdijk, N. (2003). Spontaneous speech in the Spoken Dutch Corpus. In *Proceedings of the ISCA & IEEE workshop on spontaneous speech processing and recognition (SSPR-2003)*, Tokyo, Japan.

Brady, P. T. (1968). A statistical analysis of on–off patterns in 16 conversations. *The Bell System Technical Journal, 47,* 73–91.
Bull, M. (1996). An analysis of between-speaker intervals. In *Proceedings of the Edinburgh Linguistics Department Conference'96*, Edinburgh, (pp. 18–27).
Bull, M., & Aylett, M. (1998). An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. In *Proceedings of the fifth international conference on spoken language processing* (*ICSLP '98*), Sydney, Australia, (Vol. 4, pp. 1175–1178).
Campione, E., & Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In *Proceedings of the first international conference on speech prosody* (*Speech prosody 2002*), Aix-en-Provence, France, (pp. 199–202).
Cappella, J. N. (1979). Talk–silence sequences in informal conversations I. *Human Communication Research, 6,* 130–145.
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* (pp. 346–374). Chicago: The University of Chicago Press.
The CMU Sphinx Group Open Source Speech Recognition Engines (n.d.). Retrieved 27 October, 2009, from ⟨http://cmusphinx.sourceforge.net/⟩.
Dabbs, J. M., Jr., & Ruback, R. B. (1984). Vocal patterns in male and female groups. *Personality and Social Psychology Bulletin, 10,* 518–525.
Dabbs, J. M., Jr., & Ruback, R. B. (1987). Dimensions of group processes: Amount and structure of vocal interaction. *Advances in Experimental Social Psychology, 20,* 123–169.
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.
The design of the HCRC Map Task Corpus (n.d.). Retrieved 27 October, 2009, from ⟨http://www.hcrc.ed.ac.uk/maptask/maptask-description.html⟩.
Duncan, S., Jr. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology, 23,* 283–292.
Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication, 50,* 630–645.
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 the tenth annual conference of the International Speech Communication Association* (*Interspeech 2009*), Brighton, UK (pp. 2779–2782).
Ferrand, C., Blood, G. W., & Gilbert, H. R. (1991). A continuous-flow model for phonatory reaction time. *Journal of Speech and Hearing Research, 34,* 517–525.
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 the seventh international conference on spoken language processing* (*ICSLP 2002*), Denver, USA (Vol. 3, pp. 2061–2064).
Fry, D. B. (1975). Simple reaction-times to speech and non-speech stimuli. *Cortex: A journal devoted to the study of the nervous system and behavior, 11,* 355–360.
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* (pp. 240–251). Berlin, Heidelberg, Germany: Springer.
Heldner, M., & Megyesi, B. (2003). Exploring the prosody–syntax interface in conversations. In *Proceedings of the 15th international congress of phonetic sciences* (*ICPhS 2003*), Barcelona, Spain (pp. 2501–2504).
Helgason, P. (2002). *Preaspiration in the Nordic languages: Synchronic and diachronic aspects.* Ph.D. dissertation, Department of Linguistics, Stockholm University, Stockholm, Sweden.
Helgason, P. (2006). SMTC—A Swedish Map Task Corpus. In *Working Papers 52: Proceedings from Fonetik 2006*, Lund, Sweden (pp. 57–60).
Izdebski, K., & Shipp, T. (1978). Minimal reaction times for phonatory initiation. *Journal of Speech and Hearing Research, 21,* 638–651.
Jaffe, J., & Feldstein, S. (1970). *Rhythms of dialogue.* New York, NY, USA: Academic Press.
Jefferson, G. (1984). Notes on some orderlinesses of overlap onset. In V. D'Urso, & P. Leonardi (Eds.), *Discourse analysis and natural rhetoric* (pp. 11–38). Padua, Italy: Cleup Editore.
Kendon, A. (1967). Some functions of gaze–direction in social interaction. *Acta Psychologica, 26,* 22–63.
Kousidis, S., & Dorran, D. (2009). Monitoring convergence of temporal features in spontaneous dialogue speech. In *Proceedings of the first young researchers workshop on speech technology, University College Dublin*, Dublin, Ireland.
Laskowski, K., & Shriberg, E. (2009). Modeling other talkers for improved dialog act recognition in meetings. In *Proceedings of the tenth annual conference of the International Speech Communication Association* (*Interspeech 2009*), Brighton, UK (pp. 2783–2786).
Levinson, S. C. (1983). *Pragmatics.* Cambridge: Cambridge University Press.
Liberman, M., & Pierrehumbert, J. (1984). Intonational invarance under changes in pitch range and length. In M. Aronoff, & R. T. Oehrle (Eds.), *Language sound structure* (pp. 157–233). Cambridge, MA, USA: MIT Press.
Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word, 15,* 19–44.
McInnes, F., & Attwater, D. (2004). Turn-taking and grounding in spoken telephone number transfers. *Speech Communication, 43,* 205–223.
Norwine, A. C., & Murphy, O. J. (1938). Characteristic time intervals in telephonic conversation. *The Bell System Technical Journal, 17,* 281–291.
Posner, M. I. (2005). Timing the brain: Mental chronometry as a tool in neuroscience. *PLoS Biology, 3(2),* e51.
Prieto, P., D'Imperio, M., Elordieta, G., Frota, S., & Vigário, M. (2006). Evidence for soft preplanning in tonal production: Initial scaling in romance. In *Proceedings of the third international conference on speech prosody* (*Speech prosody 2006*), Dresden, Germany.
Prieto, P., Shih, C., & Nibert, H. (1996). Pitch downtrend in Spanish. *Journal of Phonetics, 24,* 445–473.
Raux, A., & Eskenazi, M. (2009). A finite-state turn-taking model for spoken dialog systems. In *Human language technologies: The 2009 annual conference of the North American chapter of the ACL*, Boulder, CO, USA (pp. 629–637).
Roberts, F., Francis, A. L., & Morgan, M. (2006). The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation. *Speech Communication, 48,* 1079–1093.
Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50,* 696–735.
Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society, 29,* 1–63.
Sellen, A. J. (1995). Remote conversations: The effects of mediating talk with technology. *Human–Computer Interaction, 10,* 401–444.
Shipp, T., Izdebski, K., & Morrissey, P. (1984). Physiologic stages of vocal reaction time. *Journal of Speech and Hearing Research, 27,* 173–178.
Sigman, M., & Dehaene, S. (2005). Parsing a cognitive task: A characterization of the mind's bottleneck. *PLoS Biology, 3(2),* e37.
Sjölander, K., & Heldner, M. (2004). Word level precision of the NALIGN automatic segmentation algorithm. In *Proceedings Fonetik 2004*, Stockholm (pp. 116–119).
ten Bosch, L., Oostdijk, N., & Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication, 47,* 80–86.
ten Bosch, L., Oostdijk, N., & de Ruiter, J. P. (2004a). Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. In *Text, speech and dialogue* (pp. 563–570). Berlin, Heidelberg: Springer.
ten Bosch, L., Oostdijk, N., & de Ruiter, J. P. (2004b). Turn-taking in social talk dialogues: Temporal, formal, and functional aspects. In *Proceedings of the ninth conference on speech and computer* (*SPECOM 2004*), Saint-Petersburg, Russia (pp. 454–461).
Trimboli, C., & Walker, M. B. (1984). Switching pauses in cooperative and competitive conversations. *Journal of Experimental Social Psychology, 20,* 297–311.
Walker, M. B., & Trimboli, C. (1982). Smooth transitions in conversational interactions. *The Journal of Social Psychology, 117,* 305–306.
Weilhammer, K., & Rabold, S. (2003). Durational aspects in turn taking. In *Proceedings of the 15th international congress of phonetic sciences* (*ICPhS 2003*), Barcelona, Spain (pp. 2145–2148).
Wells, B., & MacFarlane, S. (1998). Prosody as an interactional resource: Turn projection and overlap. *Language and Speech, 41,* 265–294.
Wesseling, W., & van Son, R. J. J. H. (2005). Early preparation of experimentally elicited minimal responses. In L. Dybkjaer, & W. Minker (Eds.), *Proceedings of the sixth SIGdial workshop on discourse and dialogue,* Lisbon, Portugal (pp. 11–18).
Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review, 12,* 957–968.
Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society* (pp. 567–578). Chicago, IL, USA: Chicago Linguistic Society.