The University Of Sheffield.

Jan P. F. M. Gorisch

Matching across Turns in Talk-in-Interaction: The Role of Prosody and Gesture

Dissertation submitted to the University of Sheffield for the degree of Doctor of Philosophy

2012

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Jan P.F.M. Gorisch

Abstract

Understanding the design of talk-in-interaction is important in many domains, including speech technology. Although phonetic, linguistic and gestural correlates have been identified for some of the social actions that conversational participants accomplish, it is only recently that researchers have begun to take account of the immediately prior interactional context as an important factor influencing the design of a speaker's turn. The present study explores the influence of context by focussing on characteristics of short turns produced by one speaker between turns from another speaker. The hypothesis is that the speaker designs her inserted turn as a match to the prior turn when wishing to align with the previous speaker's agenda. By contrast, non-matching would display that the speaker is non-aligning, preferring instead to initiate a new action for example. Data are taken from the AMI corpus, focussing on the spontaneous talk of first-language English participants. Using sequential analysis, such short turns are classified as either aligning or non-aligning in accordance with definitions in the Conversation Analysis literature. The degree of prosodic similarity between the inserted turn and the prior speaker's turn is measured using novel acoustic techniques. The results show that aligning turns are significantly more similar to the immediately preceding turn, in terms of pitch contour, than non-aligning turns. In contrast to the prosodic-acoustic analysis, the results of the gestural analysis indicate that aligning and non-aligning are differentiated by the use of distinct gestures, rather than by the matching (or non-matching) of gestures across the adjacent turns. These results support the view that choice of pitch contour is managed locally, rather than by reference to an intonational lexicon. However, this is not the case for speakers' use of gesture. The implications of these findings for a model of talk-in-interaction are considered, along with potential applications.

Acknowledgements

Thanks to everybody who crossed my way. Without you, this would not have been possible. Thanks for being patient with me over the last years. It must have been a tough time, for Guy and Bill in particular. I hope I could give back some of this energy through and beside this work.

Part of this work was supported by the Marie Curie research training network “Sound to Sense” [grant number MRTN-CT-2006-035561]

Table of Contents

List of Figures

List of Tables

1 Introduction

The behaviour of conversational participants and their turn taking procedures have been investigated for a long time by researchers of Conversation Analysis, beginning with Sacks, Schegloff and Jefferson (1974). Because many phenomena in conversations can be illuminated by looking at the phonetics of talk-in-interaction, a strand of Conversation Analysis developed into interactional phonetic research (Kelly & Local, 1989; Local, 2003, 2007; Ogden, 2006, 2012). One focus of this strand has been the prosodic organisation of social actions performed in conversation (Ogden, 2010; Szczepek-Reed, 2012a). Similarly, the gestural organisation of social actions has been investigated (Goodwin, C., 1980; Schegloff, 1984; Streeck, 2009; Bavelas, Chovil, Lawrie, & Wade, 1992; Bavelas & Gerwing, 2007). Findings of interactional phonetics research and gesture research may be able to explain some behaviour of humans in naturally occurring conversations. One such social action is that of aligning, or showing ones understanding or empathy to a conversational partner (Stivers, 2008; Szczepek Reed, 2010a), a key characteristic of natural conversations. Such behaviour, also sometimes termed affiliation (Steensig & Drew, 2008; Müller, 1996), can for example be observed in the environment of storytelling, or in general, while conversational participants talk. At transitions from one speaker to another, it becomes relevant for the next speaker to choose either to align with the prior speaker's agenda or to move away from the prior speaker's agenda and to start something new (Szczepek Reed, 2010a; Barth-Weingarten, 2011).

One way to make dialog systems (audio and multi-modal) interact more naturally with human users would be to enable them to show understanding, empathy or affiliation. However, speech technology has major difficulties in delivering on the promise of devices which function naturally (Hawkins, 2011). Although the newest advances impress the public, e.g. the automatic speech recognition software in smartphones, the final aim of an autonomous agent is still a dream. This might be partly due to the generally accepted approach towards the interaction of user and device, which has to be designed in a way that takes the limitations of these devices into account: "There is a long road between the spoken command and its fulfilment [...]. The first step in the process is to convert the audio of speech into meaning. The two main applications of speech recognition – dictation and command recognition – have forced researchers to pursue parallel methods that balance vocabulary, accent, and context needs." (Geller, 2012, p. 14). Implementing shortcuts on that road, as can be observed in natural conversations between humans should be the aim for designers of future dialog systems. Although the importance of "context" and sequential organisation of conversation has been propagated by interactional research for a long time, speech technologists still seem to pursue a one-sided approach towards the interaction, where the incoming speech from the user needs to be decoded into an abstract "meaning".

But the conversational reality is different: engaging in social interaction implies a co-construction of action sequences. If one participant fails to produce counterparts of an action initiated by another participant, the whole interaction fails. Therefore it is required that a device correctly recognises the action of the other participant and produces a fitting response. Determining simple "meanings" of spoken utterances may lead to misunderstanding and a wrong reaction. For example the word "yeah" has one literal meaning "yes" (positive connotation), but depending on the interactional context it may work in many different, but specific ways. A "yeah" may also work like a "no" (with negative connotation) when it is used to preface a disagreement as in "yeah but". See also work on double sayings of "ja" as in "jaja" ("yeah yeah") in German (Barth-Weingarten, 2011). The analysis of the prosodic-phonetic properties is proposed to help in the distinction of different actions performed by these utterances. If the interpretation of such utterances is purely based on literal "meanings", a mis-recognition is inevitable and the device would most likely produces a non-fitting reaction. (One might argue that a basic recognition of the word "yeah" is indispensable for the recognition of the social action a "yeah" may perform.)

In real life, people *interact* with each other (see literature on conversation analysis, e.g. Heritage and Atkinson (1984) and Ten Have (1990)). In future, an interactive device is not a mere passive interpreter. Rather it actively participates in bidirectional interaction. Although word recognition and the subsequent attribution of meanings remains an important scientific aim, it is insufficient from the perspective of systems aiming at naturalistic interactions (Moore, 2007). In order to make naturalistic human-device interaction possible we should take the basic concepts into account that can be observed in naturalistic human-human interaction. Two of these basic concepts are the prosodic part and the gestural part of speech.

Interactional phonetics research

Much qualitative work has been done on the phonetics of conversation and the way conversational participants make use of phonetic and gestural resources when they talk. Starting with phonetics, the research summarised below tried to find correlates of social actions in the phonetics and prosody of the properties of the individual speakers. Such research about phonetics and especially prosody of social actions was for example done by Couper-Kuhlen (2001). High onsets and the lack of high onsets were two prosodic designs used in radio phone-in programs at points where it becomes relevant that callers say the reason for the call.

Other work on the use of the prosody of social actions is done by Kelly and Local (1989) on understanding checks. The pitch contours of word repeats are starting high in the speaker's range and quickly fall to low in order to check that a preceding utterance has been heard or understood correctly.

Work on the response token *mm* by Gardner (2001) suggests that those instances which work as continuers (like *uh-huh*) are predominantly produced with fall-rising pitch, while those which work as acknowledgment tokens are produced with falling pitch and those which work as assessments are produced with rise-falling pitch.

Specific phonetic prosodic properties of complaints have been found to have specific interactional consequences (Ogden R. , 2010). Loudness, F0 height, F0 span and voice quality can be used in cueing the relevant next action after a complaint.

Kaimaki (2011) embarked on a similar endeavour and tried to find regularities of news receipts like *oh really* and the sequential organisation of the interaction after the news receipt. Results suggest "that rising and falling pitch contours are in free variation in this interactional context." (p. 67). Free variation of pitch contours (rising and falling) was also reported for Greek data (Kaimaki, 2010).

More work on prosody in interaction was done by Ogden (2006) on the phonetic resources of second assessments. Here, the social actions agreement and disagreement "were not found to have unique phonetic properties associated with them" (p. 1772), but the use of the prosody in the prosodic context, i.e. upgrading or downgrading, could give a clue whether the agreement or disagreement was a preferred action or not. Ogden demonstrated that the adjacent turns, i.e. the first and the second assessment, are prosodically related.

> *"How a speaker making a second assessment conveys (dis-) agreement phonetically is sensitive to the other's talk, and the phonetic details of one turn are fitted to the phonetic details of the prior. Thus, even the production of talk must attend to the perception of talk, in order for the phonetic relation between the current turn and the prior turn to be such that the appropriate action is conveyed." (Ogden R. , 2006, p. 1773)*

The research summarised above tried to find correlates of social actions in the phonetics and prosody of the properties of the individual speakers. To demonstrate that specific prosodic features determine specific social actions was partly successful, but it also led to unexpected findings of prosodic dependency on the immediate prosodic context.

There is another strand of research on prosody of social actions, which hypothesises that some actions are performed by prosodically relating one's turn to the turn of the prior speaker. The paper by Ogden (2006), mentioned above, already points into this direction, that one speaker's phonetic details are sensitive to the other speaker's phonetic details.

Lerner (2002) shows simultaneous matching of prosodic features when participants co-produce turns. Such choral talk is mainly used for affiliative and cooperative actions, such as demonstrating agreement with the current speaker.

Selting (2010) analysed complaint stories and shows that the interlocutors precisely monitor each other in their displays of affectivity. Their orientation to each other is demonstrated by adapting their prosodic patterns in order to match each other.

Couper-Kuhlen (1996) looks at repetitions of the prior speaker's utterance by the immediately following one. Within the lexical repetitions, prosodic repetitions were found, which seem to be used in order to mimic or quote the prior speaker.

Müller (1996) found prosodic matching in Italian, where short tokens between speaker's turns can either be used for affiliation or disaffiliation. Those tokens which are more "in tune" and "in rhythm" with the prior speaker's utterance perform affiliation, while those which are not prosodically matching this way perform disaffiliation.

Several papers by Szczepek Reed look at the influence of prior prosodic context on the current speaker's prosodic pattern. It was found that participants orient to prosodic matching and non-matching when certain social actions are performed (Szczepek Reed, 2006). Prosodic orientation is "a practice by which participants show their awareness of other speakers' prosody in their own prosody" (Szczepek Reed, 2010a, p. 861).

A recent paper by Szczepek Reed suggests an approach to prosodic analysis which overcomes the focus on exploring the role of "individual prosodic features, speakers, locations and actions alone". Instead, prosodic analysis "must be described according to its role for both the accomplishment, and the *coordination* of actions across turns and participants." (Szczepek Reed, 2012a, p. 13).

Qualitative vs. quantitative approach

From a speech science, and especially a speech technological point of view, it seems to be worth investigating into this direction of research on prosody in interaction and particularly on prosodic matching. Future devices of dialogue systems will require to coordinate actions across turns from the user and the device. If the use of prosody by one speaker is depending on the use of prosody by other speakers and if this organisation is used by participants of naturally occurring talk in interaction in order to perform social actions, such as affiliation and disaffiliation, the same mechanisms should be adopted for dialogue systems.

Although much empirical work (see previous section) has been done investigating prosody in interaction and prosodic matching, speech technologists and designers of dialog systems tend to accept findings or claims only if the results can be based on evidence from instrumental or experimental work. As long as the claims are based on "subjective" interpretations of observations in single episodes, a generalisation is risky. Additionally, the findings remain inaccessible for the use in systems which need to work automatically on quantitative data. Therefore, prosodic matching, as it is proposed in the literature of phonetics in interaction can be seen from the speech technology perspective as being at the stage of a hypothesis. This counts for the prosodic matching hypothesis in the same way as for the hypothesis that specific prosodic features are responsible for the specific social actions. Both ask for instrumental or experimental evidence.

In the present thesis we aim to test the prosodic matching hypothesis by adopting instrumental methods which are adaptable to large data sets and create reproducible results. On the one hand this would constitute a contribution to the field of prosody in talk-in-interaction as a support of the qualitative findings on the quantitative side. It would show that it is possible to expand the qualitative analysis with quantitative methods. On the other hand it would constitute an enrichment of the field of speech technology which can adapt its algorithms in order to capture the actions which are conversationally relevant – and therefore also relevant for automatic systems. And finally it would represent a bridge between the two fields which indicates that the findings on purely qualitative analysis are trustworthy.

Advances in speech science on prosodic similarity

Not many researchers have tried to compare the prosodic patterns of adjacent turns of two speakers. But there has been some work on evaluating the similarity of prosodic contours in speech science research. For example Hermes (1989a, 1989b) compared prosodic contours of deaf speakers with model contours from a sample utterance, which should be matched.

More recently, Rilliard, Allauzen and Mareüil (2011) also compute the similarity of prosodic contours using a speech technological approach. This means that they analysed speech signals and the prosodic features with methods such as linear interpolation, the dynamic time warping technique and intensity weighting. Both studies make use of the advantage of clear speech and in Hermes' studies the compared utterances are also lexically identical, which simplifies the comparison.

Phonetic research in the past concentrated mainly on experiments using lab speech with controlled utterances and omitted naturally occurring speech. This does not only limit the scope of findings and theories to these restricted conditions, but also the methods of analysis. In this thesis we used recordings of multiparty conversations and analysed data of adjacent turns (turn transitions from one speaker to another). They are unrestricted in the lexical patterns and are therefore of highly variable acoustics and durations. This poses special requirements to the correlation of the two turns.

A study which analysed data from natural conversations for prosodic similarity is by Kousidis, Dorran, Wang, Vaughan, Cullen, Campbell, McDonnell and Coyle (2008). But they employed an averaging method of prosodic characteristics over long time frames of 10 to 60 seconds, ignoring the possibility of dependencies on a turn-by-turn basis. For this thesis it was required to develop robust algorithms, which can cope with upcoming extra challenges due to the complexity of recordings of natural conversations in general and the environment of adjacent turns in particular.

Interactional gesture research

Similar to interactional prosodic research, there is research on gestures in talk-in-interaction which examined specific gestures in relation to specific actions.

A study by Whitehead (2011) which looked at head nods in third position, i.e. after the second pair part of a sequence of actions, showed that three different types of action are performed with different gestural features of head nods. An "expansive type of nod can be used to register a prior utterance as news." A second type of nod "appears to be designed to register receipt of a prior utterance without treating it as news." And a third type of nod "embodies features of the first two types, and may be designed to register receipt and acknowledgment of "dispreferred" news." (Whitehead, 2011, p. 105)

Heath (1992) looked at nods by speakers at first position, i.e. at the first pair part of a sequence of action. It was shown that such nods are used to encourage recipients to respond in preferred ways while the utterance of the speaker is still under way. Furthermore, if the recipient produces a nod subsequently, it foreshadows their alignment with the prior speaker's utterance. In contrast, if the recipient withholds a nod, it foreshadows disalignment with the speaker's utterance.

In a paper, which discusses whether spontaneous gestures are "primarily an indication of the speech production process [...] or whether their primary purpose is communicative", McClave (2000) looked mainly at head movements. It was shown that head nods which had interactive function triggered backchannels (short responses in between talk from the main speaker). Some of these backchannels themselves can also be head nods (Yngve, 1970). However, nods seem to have more functions: "some nods also may be affirming a negative statement [...,] nods seem to be showing submission [...] and adding emphasis [...]" (McClave, 2000, p. 876)

According to Schegloff (1987), horizontal head shakes and vertical nods seem to be used differently: "a horizontal or lateral head shake can have at least three distinct uses: as a marker

or expression of the negative, of disagreement, and/or of intensification." (p.106) While the vertical nod "has a major use as a "continuer" or indicator that a recipient of speech understands that an extended unit of talk is in progress and should continue." Although such minimal gestures seem to be mere insertions into someone else's talk, they can also stand for a complete turn. For example in a short episode analysed by Schegloff, a short nod was used as "the projection of an incipient disagreement embodied in this minimal head gesture." (p. 107)

Stivers (2008) looked at head nods in the environment of storytelling. They are used by story recipients at mid-telling position in order to show their affiliation with the teller's stance toward an event reported in the story. However, whether a head nod does affiliative or disaffiliative work also depends on the position of the head nod. If the nod occurs at story completion, the story teller treats a nod as "ill fitted to that environment", as more "appropriate stance-taking responses" are required at story completion. (p. 50)

These are all studies which found evidence that specific social actions are performed with specific gestures. However, some researchers found that some social actions also depend on the gestural matchedness between speakers.

Gesture matching

In the study by Selting (2010), not only the similarity in the prosodic patterns plays a role in monitoring each others' displays of affectivity, but also the amount of gestural matching. For example by matching raised eyebrows of one speaker displaying affect by enacting raised eyebrows, a co-participant can create a "fully affiliative response" to a complaint story. (p.240) The interactional consequence is that it seems to lead to an expansion of the prior utterance. For more work on such "mutual monitoring" see work by Goodwin M. H. (1980).

Similar to Lerner's (2002) findings of choral co-productions (orally), Joh and Hosoma (2010) indicate that gestural matching can also happen simultaneously.

The interactional literature on gestural matching is weaker than the interactional literature on prosodic matching, however much work has been done on gestural mimicry in psycholinguistics and related fields.

Mc Neill (2008) observed mimicry in hand gesticulation. On the one hand such mimicry is said to be a "social interactive response", on the other hand it is used as a "tool for comprehending the other person". (p. 8)

A similar way of comprehending the other speaker was shown by Holler and Wilkin (2011a). Matching the hand and body shapes of the other speaker is said to facilitate understanding.

From an evolutionary perspective (Lakin, Jefferis, Cheng, & Chartrand, 2003), non-conscious mimicry can give an evolutionary advantage. The affiliative characteristics of mimicry is said to help fostering relationships with others. This has also been demonstrated studying capuchin monkeys which were imitated by humans (Paukner, Suomi, Visalberghi, & Ferrari, 2009). "The monkeys look longer at imitators, spend more time in proximity to imitators, and choose to interact more frequently with imitators in a token exchange task." (p. 880)

Research in mimicry has also inspired studies using a robot as conversational partner (Riek, Paul, & Robinson, 2010). Riek et al. hypothesised that a robot, which mimics the visual behaviour of people who tell it a story, is perceived more positively than a robot which doesn't.

There seems to be reason to address the gestural domain similar to the auditory domain (prosodic matching) and analyse whether social actions are also performed by matching the prior speaker gesturally.

1.1 Key questions

In the literature, it was demonstrated that for some social actions specific prosodic contours are used by participants to perform specific social actions.

It was also shown that participants orient to directly adjacent prosodic patterns by matching and non-matching. However, it was not demonstrated that the social actions performed by the matching and non-matching were doing specific, i.e. distinguishable, social actions. And a systematic sequential analysis of the acoustics is required to support the matching hypothesis instrumentally. The driving question is:

How does the choice of a prosodic contour depend on the prosodic context?

The notion of this question is supported by CA literature. Choral co-productions, which match the turns of co-participants have been investigated by (Lerner, 2002). Other authors claim that social actions depend on the prosodic similarity of adjacent turns from two different speakers (Couper-Kuhlen, 1996; Müller, 1996; Szczepek Reed, 2006; and Wichmann, 2011).

In order to answer this question, a robust measure of prosodic similarity is needed. Several limitations are imposed by the data of naturally occurring speech. It is hardly possible to control for the lexical choice of conversational participants while they speak. Additionally, it is almost impossible to find re-occurring instances of the same sequence of words. Therefore we omit the parameters which are related to the lexis (traditionally called segmental parameters). Parameters which are mainly independent of the lexical choice are prosodic parameters and are therefore our primary objective, although some interactions between lexis and prosody (microprosody), especially lexical stress may show interference effects.

Other parameters which are lexicon independent are visual gestures. In an interactional analysis it is relevant to know how co-participants use gestures in the environment of social actions. Questions related to the use of gestures are:

Do participants consistently use certain gestures to perform specific social actions?

And from a sequential analysis point of view:

Does the choice of a gesture also depend on the gestural context (is it perhaps a gestural copy)?

Gesture movements are similarly multifaceted as the acoustics of words. No trajectory of a head or hand is ever the same. Therefore it will be necessary to limit the parameters of body motion to those which can be handled by the analysis method.

1.2 Interdisciplinary approach

We adopt a combination of Conversation Analysis, and a quantitative phonetic and gestural analysis in order to give us a better understanding of how conversational interaction is managed by the participants in their face-to-face talk. This entails that the work programme is an interdisciplinary undertaking. Along the way it is demonstrated how such a combination of disciplines can be implemented.

Recent work usually treats experimental phonetic-prosodic analysis, experimental gestural analysis and interactional phonetics (using conversation analysis) as different research items. Although the three research fields can evolve and improve in their specific areas it seems to be rather difficult to relate the results and the method from one field with the results and the method from another field. In order to bridge this gap we attempt here to view phonetic-prosodic analysis, gestural analysis and conversation analysis together and take their interrelations into account and drive the interdisciplinary field of interactional phonetics further.

Traditional prosodic research tries to find relations between the prosodic shape of utterances and the communicative or "discoursal meaning" (Gussenhoven, 2004). Similarly, gestural research tries to find the relations between body movements and their communicative meaning. Following conversation analysis (in the following repeatedly abbreviated as CA), we are opposed to this search of "meanings" and instead suggest identifying social actions that are performed in naturally occurring talk and then relate them to sequence and turn organisation, the organisation of prosody (cf. to interactional phonetics research), gestures and other phenomena.

Traditional conversation analysis (CA) tries to find the regularities in the structure of social interaction. Thereby, the material used for CA has to be different to the material used in phonetic research. It has to come from naturally occurring talk rather than from clean lab-speech. On the one side, CA abstains from the cleanness of laboratory speech, where most of the variability can be controlled for, and looses statistical significance. On the other side it gains impact because it faces the reality of conversation, rather than the very limited reality of the lab. Unfortunately, because there are hardly any controllable variables in conversational speech, CA cannot revert to established phonetic techniques, which could make CA findings more robust. Additionally, the number of examples is often rather small and therefore may lack statistical power. However – and we feel this to be important – much insight can be gained from a qualitative analysis, even if only a few samples of conversational data are available. This is not to be received from a mere quantitative phonetic approach. Despite a poor statistical power we do have opportunities to improve our knowledge about the regularities in the organisation of social actions and their relevance for a functioning conversation:

> *"One point which seems increasingly clear is that, in a great many respects, social action done through talk is organized and orderly not, or not only, as a matter of rule or as a statistical regularity, but on a case by case, action by action, basis." (Schegloff, 1987, p. 102)*

See Sacks, Schegloff and Jefferson (1974) for more arguments why such qualitative analyses are most of the time advantageous to pure quantitative analyses.

When we look at traditional phonetics research, the known theories and the usual testing methods limit the scope of analysis to narrow laboratory settings. Therefore they may not satisfactorily be regarded as conclusive for the real-world setting of naturally occurring talk. The latter is required for conversation analysis. "The key point is that, although CA's methodology may be applied to interactions in certain kinds of experiments, interviews, or simulations and fictional constructs, basic research in CA uses only naturally occurring interactions as data." (Drew, 2004, p. 78). This is because conversation is the most basic use of human language and because the interactional participants (our subjects) are directly involved in the production process of social actions. This has also found its way to researchers of neuroscience (Scott, McGettigan, & Eisner, 2009) who argue "that there is a central role for motor representations and processes in conversation [...], an essential aspect of human language in its most basic use." (p. 301)

The approach described in this thesis is an attempt to narrow the gaps between the before mentioned research fields. On the one hand, we analyse the interactions qualitatively. This helps us to find the sequence of elements which organize the talk, i.e. the objectively gained categories which are based on social interactions – objective, because the categories are based on the apparent and openly observable orientation of the participants towards them. On the other hand, the phonetic analysis highlights the details of the performed talk, similar to previous work in interactional phonetics, and additionally provides us with objective quantities of the details. Thereby we make both objective, the categories with their interactional characteristics and the quantitative parameters with their acoustic-gestural quantities.

The focus of the phonetic analysis lies on the spoken part of verbal utterances and generally ignores the gestural part. If visual analyses are performed in phonetics research it is limited to the visible part of articulation. Introducing the gestural part may help to balance the modalities which are used by the conversational participants. The combination of the two methods (qualitative and quantitative) and the combination of the two modalities (audio and visual), as we propose, opens new perspectives for the analysis of face-to-face talk. Regarding phonetics, rather than analysing de-contextualised individual speech units (such as tones, phones, phonemes, words, intonation phrases, etc.) our aim is to analyse the phonetics of speech units in their sequential adjacency, their phonetic context, their gestural context and with respect to their interactional performance. This means that we investigate the social actions which are performed by participants depending on a certain phonetic context. In other words, rather than analysing prosodic features solely with reference to communicative functions, we analyse them

with reference to the prosodic, gestural and lexical context and with reference to social actions, the latter as performed during CA.

CA is criticised in several ways from the ethno-methodological perspective, which has been discussed by Ten Have (1990) who has addressed most of the criticism. One critique which still seems to be justified from the social psychologist's point of view is that CA literature is usually based on discussions and findings from one single interpreter who in most cases is the author. CA argues that the researcher's findings are grounded in the recording and confirmed by the participants' behaviour. The objectivity of the analysis can even be increased by introducing further judges – although it is not common practice in laboratory phonetics to let the analysis of spectrograms being checked by other "judges" (which probably should).

Other critique can be put forward from the perspective of speech technology, which might be interested in adapting relevant rules, techniques, procedures, methods and maxims used by participants in order to achieve interactional goals. Such rules etc. are proposed in the CA literature (Sacks, 1984). However, if it is intended to adapt one such rule or another in an automatic dialogue system, the relevance of this rule needs to be statistically tested. This is usually done with data collections large enough to achieve statistical significance, which is not the case for most studies in CA as it is not required for CA. In order to overcome this dilemma and in order to make CA results more attractive for other disciplines like speech technology, it seems advisable to statistically test the relevance of the rules.

Another weakness which may be attributed to CA is the subjectivity when interpreting the physical counterparts of the collected instances of contrasting categories, e.g. by-ear identification of pitch contours which are associated with social actions. F0 contours or other visual representations of prosodic features usually support the impressionistic record-making and transcribing of audio recordings, but it is not tested or verified, whether the perceived pitch is mapped equally well by the visual representation. In order to distinguish certain interactional categories which can be found in the corpus on the basis of phonetic-prosodic measures in an objective manner we decided to develop an automatic procedure based on the acoustic signal. Visual representations of prosodic parameters are merely used for illustrative purposes, but not as the basis for analysis.

Although conversation analysis has been criticised in some points, the qualitative methodology has the main advantage of analysing the sequential organisation of talk-in-interaction. Therefore it opens up the opportunity to analyse the sequential use of acoustic features, rather than to analyse the acoustic features in the traditional de-contextualised way.

Natural conversation is most of the time performed face-to-face. This implies that besides the verbal aspects, such as the participants' wording, the manner of how they say it and the way of how they use acoustic features sequentially, the non-verbal aspects, such as gesticulation and facial expressions play a major role in performing social actions, too. While some social actions may be performed with individual gestures (for example with emblems such as the "okay" sign (McNeill, 1992)), other social actions may be performed with gestures in sequential use. For example gesture mimicry is observed when a hand gesture from one speaker is copied by the other speaker. Gesture mimicry has been demonstrated to be used to demonstrate "a mutually shared understanding of referring expressions" (Holler & Wilkin, 2011a).

Most communication is still happening face to face and visual communication for some social actions is considered to be a crucial factor in conversation (Bavelas, Chovil, Lawrie, & Wade, 1992), (Bavelas & Gerwing, 2007). The visual aspect deserves the respective amount of attention and is therefore included in this thesis.

In order to find out whether specific gestures are used by participants to perform specific social actions we correlate the gestures which accompany the verbal utterances with the social actions on the one side. On the other side, in order to find out whether a gestural match or non-match is responsible for specific social actions we correlate the adjacent gestures from both speakers.

Overall, we adopt a combination of two approaches: a multimodal and an interdisciplinary approach, which engages with a holistic perspective of social interaction in naturally occurring conversation.

1.3 Structure of the thesis

The three main topics prosody, gesture and conversation analysis in interactional phonetics are introduced in Chapter 2. First we briefly review the conversation analytic research methodology and clarify definitions for social action and context. The basic conversational resources, such as prosody and gesture are discussed.

The material and method are introduced in Chapter 3. The transcription and gesture annotation of the data for the purpose of conversation analysis are described. Conversation analysis is used in Chapter 4 to evaluate the interactional regularities of alignments and non-alignments in adjacent turn sequences. As argued above, in order to increase the objectivity of the outcomes of the conversation analysis stage, emphasis is put on the reliability of the details found in the sequential organisation of alignments and non-alignments. An inter-rater agreement test is described in Section 4.3.

The two chapters 5 and 6 are connected to the interactional analysis according to the different modalities. In these two chapters we look for relationships between interactional categories and the acoustic and gestural properties.

The acoustic analysis in Chapter 5 is used to investigate the prosodic properties which can be encountered in the environment of social actions. Prosodic similarity is suggested to be one driving factor in the distinction of two basic social actions: alignment and non-alignment. Two similarity metrics are reported, which are evaluated according to artificial prosodic contours and applied to the real examples which have been collected at the interactional categorisation stage (Section 4.2).

Chapter 6 is split into two parts. The first part (Section 6.1) comprises the gestural analysis and is used to investigate the non-verbal properties in the environment of the specific social actions. Two investigations were conducted. First, it was investigated how the individual gestures and facial displays of *one* participant are related to the social actions. Second, it was investigated how the relation of gestures from *both* participants are related to the social actions. The second part (Section 6.2) combines the prosodic modality and the results from the acoustic analysis from Chapter 5 with the non-verbal modality and the gesture data.

In the final Chapter 7 we discuss the implications of our attempt to model prosodic matching and the use of gestures in relation to the basic social actions of alignment and non-alignment. With respect to these social actions in naturally occurring conversations, the acoustic analysis of prosodic similarity is among the first ones with this idea being reported. The findings are discussed with particular reference to the needs of dialogue systems. However, further implications are explored, for example for teaching prosody to second language learners, research in developmental disorders such as autistic spectrum disorder, the use of gestures in conversation of people with speech disorders such as aphasia. Lastly the findings are also related to the growing research on empathy.

2 Prosody, gesture and talk-in-interaction

Prosody is an inevitable accompaniment of talk and in face-to-face dialogue gestural cues are inevitably visible to conversational participants. Both have the potential to influence the participants' behaviour. In this chapter, different viewpoints on these phenomena are presented with a special emphasis on research in conversation analysis and interactional phonetics.

2.1 Background on conversation analysis

In this section, some basic terms used in the conversation analytic literature are explained and some key concepts that are characteristic for the CA methodology are discussed. Their usefulness for the current study is evaluated.

2.1.1 Social action and context

The terms *social action* and *context* were already used in the introduction, however no definition of these terms was given there. Drew (2004) explains what conversation analysts understand by social actions:

> *"... conversation is not, to adapt Wittgenstein's phrase, "language idling." We are doing things, such as inviting someone over, asking them to do a favor or a service, blaming or criticizing them, greeting them or trying to get on first-name terms with them, disagreeing or arguing with them, advising or warning them, apologizing for something one did or said, complaining about one's treatment, sympathizing, offering to help, and the like. These and other such activities are the primary forms of social action, as real, concrete, consequential, and as fundamental as any other form of conduct. That such actions as these – for instance, inviting, complaining, and disagreeing – are at the heart of how we manage our social relationships and affairs hardly needs explanation. So when we study conversation, we are investigating the actions and activities through which social life is conducted." (Drew, 2004, p. 74 f.).*

The orientation of conversational participants towards each other manifests the performed actions. A complaint is not a complaint if no complainable can be identified, or if the co-participants don't treat a complaint as a complaint (unless a misunderstanding occurs, which is subsequently repaired). This is what Drew means by "consequential". The surrounding of a complaint makes a complaint recognisable as a complaint for the analyst. The aim of Conversation Analysis is to find regularities in the sequential organisation of talk which can then be used in order to explain specific social actions.

Further, CA is fundamentally concerned with "context". Two notions of this term were presented by Mandelbaum (1990): talk extrinsic and talk intrinsic context (p. 337). While talk-extrinsic context is determined by characteristics including age, status, and gender of participants and their socio-cultural background, talk-intrinsic context is constituted in the local details of the ongoing talk. Many of the social actions are "next-positioned", which means that they respond to something immediately prior. Work in CA (Pomerantz, 1978 and Drew, 1987; summarised by Mandelbaum, 1990) shows that social actions can be explained on the basis of that *local* sense of context. Such local detail builds up the context for specific conversational actions: "conversation analysts ask about the social actions conducted in and through talk. Context for conversation analysts is built in and through talk." (Mandelbaum, 1990, p. p. 346).

This means that social actions and eventually some cultural norms can only be discovered through the analysis of the context, i.e. the talk that contains the orientation of the participants towards such systems.

2.1.2 Alignment and affiliation

The focus of the current thesis is on the social action of alignment. Barth-Weingarten (2011) defines the terms (dis)alignment in contrast to (dis)affiliation: *(dis)alignment "*is used as a purely structural notion, referring to the (lack of) endorsement of the sequence/activity in progress, and thus contrasts with the notion of *(dis)affiliation*, which is understood as a (lack of) endorsement of the previous speaker's evaluative positioning, or stance." (p. 161). This definition originates from Stivers (2008), who has used the term "aligning" to describe actions by a second speaker which support the activity being undertaken by the first speaker. She illustrates this from storytelling, showing that a token such as "uh huh" produced by a new speaker "supports the structural asymmetry of the storytelling activity" (p. 34). This type of alignment also accords with Schegloff's use of the term in describing participants' behaviour in telephone closings: "the recipient can then elect to introduce some new sequence or topic, or can align with the caller's preparedness to proceed to the closing of the conversation; this way of proceeding is, then, designed to be consensual" (Schegloff, 2007, p. 257).

On the other hand, "competing for the floor or failing to treat a story as either in progress or – at story completion – as over" is "disaligning" (Stivers, 2008, p. 34). One case of disaligning with the telling activity is a "mid-telling initiation of a sequence [which] disrupts the progressivity of [a] telling, and thus [the] response is analyzable as obstructive rather than facilitative." (p. 35). A similar phenomenon is described by Steensig and Drew (2008) in their review of different types of questions: "Asking a question is not an innocent thing to do. Often questions challenge or oppose something a co-participant has said or done, thereby creating possible interactional disaffiliation." (Steensig & Drew, 2008, p. 7). Steensig and Larsen (2008, p. 126) show that part of what is involved is that the question is a "disaligning move". Drew (1997) indicates that repair initiations too can have this property: "Matters of comprehension and repair shade into matters of accord or (mis)alignment between speakers." (Drew, 1997, p. 72). Thus it appears that dis- (or mis-) alignment can be accomplished by different types of action including questions and repair initiations, as well as more obvious incursions into the current speaker's turn (French & Local, 1983); (Kurtić, Brown, & Wells, 2009) or into the current speaker's story in progress (Stivers, 2008). Since the interactional interpretations given to prosodic matching in the prosodic literature described earlier (see interactional phonetics research of Müller (1996), Szczepek Reed (2012) in the introduction) are close to Stivers' and Barth-Weingarten's conceptualisation of "alignment", we employ this term in the present study. We further use "non-alignment" to indicate the absence of (positive) alignment with the prior turn. This is meant to be a neutral cover term which includes cases of disalignment or misalignment as described in earlier research.

The definition of alignment and nonalignment can be further expanded. There are several ways in which talkers may not be aligned. They range from disagreement over disaffiliation to a simple lack of matching a first pair part of a social action with the expected second pair part. According to the "preference organisation" of talk, for example, a question as first action is expected to be responded by the recipient as paired second action. The preferred second action after a greeting is a return greeting, a request should be either granted or rejected. Failing to match the first pair part with a matching second pair part is considered as a dispreferred action. It can also be considered as a non-aligning action.

Drew and Walker (2009) looked at complaints that were performed on behalf of another participant. Against their expectations that participants on whose behalf was complained would always show affiliation with this complaint, such participants did even show disaffiliation. For example, they express that they would not go "too far in escalating the complaint to a point beyond which the 'complainant' is comfortable." (Drew & Walker, 2009, p. 2400). The findings reported show that both affiliation and disaffiliation to a complaint launched by another participant on one's behalf are regulative resources to express how far someone else may go in

complaining. A disaffiliating move may then be seen as a move away from the continued action of escalating a complaint.

Support for a distinction into the two actions of alignment and non-alignment can also be found in an article by Wells (2010) on a child carer-interaction where the child orients to the direct context by displaying "alignment with an ongoing activity [or] initiating a new action or sequence" (p.244). Szczepek Reed (2012) also argues for a "fundamental conversational activity of displaying the sequential status of an immediately next turn either as a continuation of prior talk, or as the beginning of a new project." (p. 14).

These definitions help to clarify the concepts of alignment and non-alignment. There is the notion of "continuation of prior talk", "endorsement of the activity in progress" or "alignment with an ongoing activity" for alignments and "the beginning of a new project", the "lack of endorsement of the activity in progress", support for "a structural asymmetry of the storytelling activity" or initiation "of a new sequence or action" for non-alignments. The question then arises about how one can, for instance find out whether prior talk is continued or whether a new project has begun.

According to CA, answers to such questions can be found by analysing structural elements in the turn sequence. Such structural elements are (i) the prior turn (ii) the second / target turn (iii) the subsequent turn and possibly (iv) the sub-subsequent turn. Additional to the verbal part is the gestural part and within both modalities there is the overlap across speakers and stretches of silence which are either filled with gestures or not. The analytical challenges are that all these phenomena are intertwined and depend on each other.

In order to disentangle these dependencies of actions and turn organisation, it is important to have an approach to the description of turns at talk in principal. The next section addresses this topic.

2.1.3 Turn construction

Doing conversation is using practices to co-construct talk (Jacoby & Ochs, 1995). According to Jacoby and Ochs (1995), the research aim of CA is to find out these practices through identifying patterns in that talk and patterns in the way in which it is co-constructed.

An example (Extract 1) is used here to introduce some concepts of CA that are important for the present study: Turns at talk and turn taking, turn design, social action, and sequence organisation (Drew, 2004). This example also illustrates the types of actions that we are going to analyse in more detail in the remainder of this thesis. In the face-to-face conversation of Extract 1, three participants are discussing recent developments in a shared research project. Only two participants (A and B) are speaking in this extract, however the third speaker (C) is also engaged in the conversation as can be seen from the visual layer of the transcript. The corpus from which it is taken is described in detail in Section 3.1.1 and the transcription conventions in Section 2.1.5 and in Appendix A.

Extract 1: "output format" (C-04:40)

```
        --------------------((nod, blink))------------------
1 C: [                                                    ]
        gaze down------,,,--gaze to C---((gesticulation with right hand directed to C))
2 A: [u:m in the output format for the for the] task you know so
        ---,,,-------gaze down-----------
3    what what you're what you'
        -------((nod))------
4 C: [                  ]
        ---------,,,-------mid gaze---------------,,,-------
5 A: [re gonna do the] analysis on .hhh um so
        ---gaze to C-----,,,----------gaze down----------------,,,--------------gaze to B---
6    making sure that .h I can take (0.2) get that information out of
        ------------((gesticulation))----------((gesticulation hold for a moment))
7    the GDF as it st* as a state of the moment .hhh
        ((nod, blink))
8 B: [uh huh]
```

```
       ---------------------------------((gesticulation))----------------------------------
9 A: [and if] I find something that you can't get out of it then I can
       --((end gesticulation))-
10    add that to the add that to the GDF format (0.7)
```

The concept of turns at talk and turn taking includes that "one speaker takes a turn and is followed by another speaker [taking a turn]." (Drew, 2004, p. 80). A turn can be a single word such as the "uh huh" by Participant B in line 8, or it can be long as Participant A's lines 2 to 7, who makes an extended report. The visual modality complicates things, but only slightly: Participant C's gestures in line 1 and 4 can also be seen as turns, although no speech is produced simultaneously (Sacks, Schegloff & Jefferson, 1974; Drew, 2004). They are just produced in overlap with Participant A's speech (indicated with square brackets "[...]").

Words, phrases, sentences, gestures are building blocks of conversation from which turns are made. They are also called turn constructional units (TCUs). What a turn is and how long it is – i.e. how many TCUs it may contain – depends on the interactional work it is designed to do. "First, a speaker selects what action the turn will be designed to perform. Second, he or she selects the details of the verbal constructions through which that action is to be accomplished." (Drew, 2004, p. 82 f.). It can for example be designed to select a next speaker, or to make a report, or to invite someone continue reporting.

The notion of turn constructional units (or units in general) is under debate (see Section 2.3) and a demarcation of such units can be problematic, as an attempt of the current example from Extract 1 shows. Following Drew, the whole extract from line 1 to line 10 can be seen as a single turn from Participant A, as it is designed for the action of extended reporting on a specific task.

In the lines 1-7, neither of the other participants speaks in the course of the turn, though Participant C uses nonverbal gestures. There are several places where the speaker makes restarts, repeats words or parts of words, takes inbreaths, makes short pauses, but some techniques are employed by the current speaker to prevent other speakers to take up the floor at such points. One common practice is for example to avoid breathing at points where a syntactic construction comes to a point where it is complete, and continue speaking until a place where the grammatical construction is incomplete and breathe there. An example is here the "so making sure that .h" in line 6. Such "rush-throughs" through transition relevance places have been demonstrated by Schegloff (1982). Also, the "so" at the end of line 2 and continuous hand gesticulation projects continuation in this case. The gesture hold (end of line 7) occurs during an extensive inbreath and after the grammatical construction comes to a point of completeness. Exactly at that point, Participant B produces verbally a short turn ("uh huh", line 8). It is suggested that this happens because Participant A is gazing towards Participant B. Such gazing at potential transition relevance places is a way in which a current speaker may select a next speaker (Lerner (2003), Walker (2012)). Participant B's "uh huh" accompanied by a nod and a blink (line 4) is designed as continuer (Schegloff, 1982), handing the floor directly back to the prior speaker, namely Participant A, who continues his reporting with hand gesticulation.

After several chunks of TCUs have been passed, in line 7, Participant A selected Participant B by gazing at her. She takes up the possibility to speak, producing the short turn "uh huh". The work it does in the conversation is to invite Participant A to continue speaking. Participant B selected this social action from a number of other possible actions or activities to perform in that turn. She could have chosen to stay silent, or to produce an extended turn herself, or to insert a short utterance introducing a change of direction for the talk. Such an instance can be seen in Extract 2.

Extract 2: "block booked" (C-16:57)

```
       gaze to C
1 B: yeah
       gaze to B,,,,---------gaze to A--------,,,,,,,,,,,,,,,,,,,,,,,,-----gaze to B---
2 C: .hhhh u:hm uh the lab has been block booked again (0.5) .hh
```

```
       ((blink))
3 B: by::
       ,,,,----mid gaze--((body movement))--,,,----gaze to B----((blink))-----------
4 C: .h (alveolar click) Jules (0.5) uh from: u::h psychology hh
       mid gaze(head shake, blink, raised eyebrows))-,,gaze to C
5 B: um  (0.7)  I  don't  know  this  person
        ((nod))
6 C: [      ]
       ------,,
7 B: [(0.5)]
       gaze down
8 B: so
       ----((nod))----
9 C: [            ]
       gaze to C,,---gaze down--,,-----gaze to C----------
10B: [(0.5) (0.5)] they (0.5) (0.5) well ...
```

Participant C reports that a laboratory "has been block booked again", which prevents the researchers using it for their purposes. It can be interpreted as an accusation, however the accusation is not directed at anyone specifically, as the turn contains an agentless verb ("has been block booked"). Following the short turn "by::", Participant C continues to talk, but his turn "Jules", prefaced with a short inbreath and an alveolar click, is only understandable as a second pair part: it is responsive to "by::" which is an "increment initiator" (Lerner, 2004, p. 154). Thus B's turn serves to change the direction of the talk, moving from Participant C's agenda, which is a reporting of an event which may cause trouble to some task of the researchers, to Participant B's agenda, which is a specific request for clarification of the details of that event. Participant B's choice of this action is substantially different to the choice of action in Extract 1, where the prior Participant A continued with his own agenda after B's short turn. There, Participant B also had the possibility to change the direction of the talk in the way that C does in line 4 of Extract 2. This could for example be done by challenging the prior speaker's statement or by asking for clarification.

> *"So we are studying the use of language in conversation (turn design) employed to do things in the social world, and we focus on the social responsiveness of the sequential organization of these activities being conducted in conversation." (Drew, 2004, p. 86).*

To come to a conclusion of what the social actions are, Schegloff recommends grounding the analysis in the understanding of that social action by the participants. This means that we can find an "empirically grounded account of action" (Schegloff, 1996, p. 172), when participants demonstrate their understanding of the other speaker's action. In Extract 2, Participant C demonstrates that he understood Participant B's action in line 3, i.e. her initiation of a move away from his agenda and a start of a new agenda. C does exactly this in line 4 by providing an increment in that new course of action. In Extract 1, Participant A demonstrates his understanding of Participant B's action of handing back the floor, by continuing on his own agenda and not treating B's turn as an invitation to shift the topic. Thereby the participants themselves gave the answer to the question what the social actions are. And they do this by creating and orienting to the immediate context. The task of the analyst is merely to disentangle the turn-by-turn, unit-by-unit development of the talk.

Extract 1 and Extract 2 display sequences of actions. In (1), the sequence consists of three turns or slots, (i) an extended report, (ii) a continuer, (iii) a continuation of the prior speaker's report. In (2), the sequence equally consists of three turns, (i) a report, (ii) a short turn that diverts the course of talk away from the prior speaker's reporting, (iii) an account from the prior speaker (C) who follows the new course of action.

2.1.4 Repair

An important method to analyse actions in talk-in-interaction is the evaluation of the participants' orientation to the utterances of each other, when a trouble arises in the conversation. Such situations are usually recognisable by some sort of repair.

In Extract 1, Participant A is not just reporting on his research activities. He is also displaying how these activities and their results impact on the other participants of the meeting. His task is to make some information available to Participant C, because "the output format" is something "what you're [Participant C] gonna do the analysis on". He adds "so making sure that I can take (0.2) get that information out of ..." with a self-repair substituting the verb "get" for "take". This little change can have substantial consequences on the interpretation of that stretch of talk by the coparticipants. The verb 'take' requires a more active behaviour in order to fulfil that task. It might even involve a further task of, for example storing the information for Participant C in some place. The verb 'get' describes a more passive behaviour and may signify that, if Participant A can 'get' some information out of something, it is very likely that Participant C can also 'get' it the same way. "Through these self-repairs the speaker adjusts the design of the turn so as to better fit the turn's project and more effectively to convey what [he] means to." (Drew, 2004, p. 96).

If some turn design which may cause trouble for the recipient is not repaired, the misunderstanding can become evident through the subsequent turn of the recipient, not indicating the trouble directly, but indirectly. Then the first speaker can choose to self-correct in the directly following turn. The misunderstanding can also be addressed by the recipient directly after the problematic turn. Such examples of repair and more examples of self-repair or other-initiated self-repair can be found in Drew (2004, p. 94 ff.).

The first two extracts used here to introduce the basic concepts of CA research have addressed rather complicated issues in the analysis of social actions, but these examples have been chosen on purpose: They exemplify sequences in which short utterances (turns), inserted between turns of another speaker, performed specific social actions. One utterance was the "uh huh" in Extract 1, apparently used to align with the prior speaker's turn. The "by::" in Extract 2 is another example of a short utterance that is inserted between turns of another speaker, however doing different interactional work than the "uh huh" in Extract 1. The "by::" seems to disalign with the prior speaker's turn.

In order to investigate such basic social actions, CA draws upon interactional regularities such as repair, as Drew has put it:

> *"The aim of CA is to identify precisely those methods, procedures or practices that enable participants to construct their talk to do, and to be recognized as doing, what they mean to be doing." (Drew, 2004, p. 94).*

2.1.5 Transcription of talk-in-interaction

The basic tool for analysing conversation is a transcript which contains as much information as possible that is available from the recordings and "as few *a priori* decisions about the relevance of particular phonetic events as possible" should be made (Local, Kelly, & Wells, 1986, p. 413).

One main requirement is to "characterise the sequentially organised properties of the talk." (Ball & Local, 1996, p. 71). Instead of transcribing sentences, each 'turn' at its actual position (e.g. in overlap) is to be transcribed, as "a turn's talk will be heard as directed to a prior turn's talk, unless special techniques are used to locate some other talk to which it is directed" (Sacks, Schegloff, & Jefferson, 1974, p. 728). None of the features that might be relevant to the interactional participants and the organisation of their talk should be omitted. Such features can be:

- "noticeably louder portions of utterances" (Ball & Local, 1996, p. 73) indicated by capitals,
- "lengthened segments" indicated by colons ":",
- "silent intervals" are timed and indicated, e.g. in seconds "(0.3)",
- "word fragments, non-word vocalisations such as inbreaths and obscure parts of utterances are indicated" (p. 74),

- "prosodic parameters [...] (i) above the orthographic text: relative pitch height and on-syllable pitch movement represented impressionistically within staves which designate the limits of the speaker's normal pitch range; (ii) below the text: IPA symbols and extensions" (p. 83).

More details that are potentially relevant for conversational participants are: timing and placement of speech (e.g. overlap), sound qualities (such as sound stretching, emphasis, loudness, marked pitch changes, and certain intonational features), in- and out-breaths, laughter, cutoff words or sounds (Drew, 2004, p. 78).

Local, Kelly and Wells (1986) suggest "that the transcription of pitch, loudness, tempo, rhythmic and other phonic phenomena be carried out consistently and impressionistically ..." (p. 413).

In face-to-face conversations additionally to the features mentioned above, also the visual properties of the interaction are likely to be relevant for the development of the conversation. Goodwin C. (1980) suggests gesture/gaze layers above each line containing the non-vocal aspects of spoken language, one for the current speaker and one for the principal recipient. Gestures contain head, manual and body gestures indicated by orthographic descriptions in round brackets. When the gestures start and end is indicated by the dashed line. The direction of gaze is also described orthographically and with commas indicating when gaze shifts. These are relatively simple transcription rules compared with the inventory of detailed symbols illustrating the shape of eye-brows, hands, hand shapes, as described by Birdwhistell (1970). They are detailed enough and practical for CA type analysis.

> *"The significance or relevance of these details may not be (probably is not) apparent when one is transcribing the recording of an interaction; they may come to have any significance only as one begins to analyze the data. But at the time the transcription is made, all that lies ahead; the transcriber attempts only to capture on the page, as faithfully as possible, in as much detail as possible what was said and how and when it was said." (Drew, 2004, p. 78).*

This premise holds for all modalities that are available to the participants – and to the analyst through the transcripts. However, Kelly and Local (1989) argue that even the phonetic alphabet of the IPA "is likely to obscure the kinds of dynamisms, continuities and overall phonetic patternings [because of the] non-systematic, cross parametric nature of [that system]" (Kelly & Local, 1989, p. 113). Their aim was to come to a phonological account for conversational speech. Therefore, they decided to look for "features that are suspected candidates for long-domain relevance [they] call stretches in the interpretative phase of [their] work. […] the establishment of stretches is the logically prior activity" (p. 118). They continue: "It is a common, but fallicious, belief that some features are distinctive and some 'non-distinctive', and that these 'non-distinctive' features are redundant to analyse. […] some features which may be claimed to be redundant in the analysis of lexis may well prove to be distinctive when considered in the context of grammatical or interactional analysis or more extensive contexts in general." (p. 118). This suggests that although it is desirable to focus on every phonetic detail in order to make impressionistic transcription, it is practically impossible. The question arises whether any transcribed representation of the recording can ever be sufficient at all. But it seems to be possible to transcribe the conversation at some level of detail, e.g. at the word level, and to make the acoustic analysis in a subsequent step, which would otherwise be done by the transcriber. As long as the argumentation of the social actions can be based on what *is* transcribed, the acoustic analysis can be used to support it and to derive the acoustic cues that are used by the participants.

2.2 Resources for conversation

Much work has been done on the raw materials or resources which help participants to construct turns. Apart from the lexical and syntactic choices for the construction of a turn, researchers in

interactional phonetics and CA found that prosody and gestures are further such resources which are employed by participants in order to construct turns, organise turn transitions and perform specific social actions. In the review of such literature we begin with the prosodic ones.

2.2.1 Prosodic resources

In the design of certain actions, prosodic features play a particularly important role. In this section, examples are given of studies where a range of prosodic features have been shown to accomplish different kinds of interactional work. Along the lines of the traditional prosodic analysis where specific communicative functions and meanings were associated with specific prosodic characteristics, interactional phonetics also analysed specific social actions and tried to find specific prosodic-acoustic correlates. This was partially successful, as the findings of the following articles demonstrate.

2.2.1.1 Specific actions – specific prosody

A study which looked at the prosodic characteristics of single turns which performed different specific social actions was done by Couper-Kuhlen (2001). In the conversational setting of radio phone-in programmes, after a greeting sequence, the conversation comes to an anchor position where the caller can either give the reason for the call or merely project this action or insert something immediate or urgent. Two contrasting prosodic designs are used to cue the status of the talk at anchor position: high onset vs. absence of high onset. High pitch on the first stressed syllable is used to carry out the action of saying the reason for the call, while the absence of such high onsets indicates something other than saying the reason for the call. The choice of using these prosodic cues is independent of the wording employed by the caller. Similar utterances can be produced with different prosodic design, determining the caller's turn as either a part of a multi-unit turn that constitutes the action of stating the reason for the call (with high onset), or as a preliminary to the reason (without high onset). Such a shift from a "normal" pitch range to a higher register can be interpreted as a register non-match from one turn to another. As Couper-Kuhlen demonstrated, such a non-match can perform a specific action. Here, it is the move by one participant from preliminary small-talk to saying the reason for the call.

In a study by Kelly and Local (1989) on understanding checks, an interviewer in a sociolinguistic survey used test words of the dialect in question in order to find out if the interviewee knew them or used them. After the interviewer said the test word, the interviewee sometimes repeated this test word. On some occasions the repetitions were produced as understanding checks, which "are designed to check that a preceding utterance has been heard or understood correctly and they require a response before proceeding with any further talk." (p. 272). One of the phonetic correlates of these understanding checks is a falling pitch contour, starting high in the speaker's range and quickly falling to low. It is also interesting to note that the "whole word is [not only] louder than the interviewee's usual talk [... but also] louder than the interviewer's preceding turn." (p. 279): thus in part at least, the understanding check is designed phonetically with reference to the interviewer's prior turn. From the perspective of "matching" or "non-matching", it can be concluded that the lexical repetitions do not match the intensity of the prior speaker. If this phenomenon can be supported by further evidence, it would suggest that understanding checks are performed by lexically repeating the prior speaker's utterance while not matching the prior speaker's intensity.

Other prosodic features have been shown to play an important role in the design of complaints. Besides lexical and sequential properties, two types of complaints can be distinguished by their phonetics (Ogden R. , 2010). One type of complaints "make affiliation a relevant next action, thus proposing a continuation of the complaint sequence; while with [the other type of complaints] a complainant can propose closure of their own complaint sequence." (p. 99) The phonetics of the first type are: loud, high F0 above the speaker's average, F0 peaks high in the speaker's range and wide pitch span. The phonetics of the second type are: relatively quiet, low F0 in the speaker's range, overall "lax" setting (e.g. creaky voice quality) and narrow pitch span.

For the environment of turn transitions or places where turn transitions from one speaker to another become relevant, Selting (2000) also states that even contrasting contour shapes can perform similar social actions: "In general, rising and falling pitch at the end of possible syntactically complete constructions can be used to signal turn-yielding", but she still makes a difference between those specific prosodic patterns (rising and falling) and other specific prosodic patterns: "final level or only slightly rising pitch is used to signal turn-holding for more to come" (Selting 2000, p. 510). Prosodic contours which at first sight seem to be contrasting (here, falling and rising) are bundled and contrast interactionally with other contours, which also seem to be contrasting at first sight (here, level and slightly rising).

Prosodic features are not always systematic in relation to the design of turns or actions. Kaimaki (2011) analysed non-valenced news receipts like "oh really", and valenced news receipts including "oh good", "oh how great". She tried to find regularities in the phonetic organisation of the news receipts and the sequential organisation of the interaction after the news receipts, like confirmation, confirmation+elaboration, continuation or more talk by the news recipient. Although the frequencies of these uptakes seem to differ between rising and falling "oh really", indicating intonational "contrasts", results suggest "that rising and falling pitch contours are in free variation in this interactional context." (p. 67).

For Greek data, similar variation of pitch contours were reported (Kaimaki, 2010). The analysis "of *response-to-summons turns* suggests that the choice of falling or rising tune does not appear to have consequences for the design or subsequent development of the talk. Nor is there evidence in the interactional behaviour of the participants that the choice conveys a difference in pragmatic nuance." (p. 213)

The interactional phonetics literature shows that specific social actions are sometimes related to specific prosodic features. Sometimes, no specific prosodic features could be identified for specific social actions.

2.2.1.2 Response tokens

The phenomenon often referred to as "backchannel feedback" is another type of action like complaints, understanding checks etc., which has been investigated by other research fields than CA. They are particularly relevant to the work that is reported in this thesis as they occur frequently in conversational talk. Their sequential use and the interactional work they do in specific points of a conversation have been analysed extensively and their prosodic properties have been investigated.

Backchannels

There are many terms for such short utterances as "uh huh", "mm hmm", "yeah", "okay", "mm", etc. in the literature. Ward and Tsukahara (2000) call them "back-channel feedback" with "infinitely many non-lexical variations" (p. 1183), as an "extended category" for "the short utterances produced by one participant in a conversation while another is talking." (Ward & Tsukahara, 2000, p. 1177). "They also seem to comprise many of the phenomena that have been studied as 'listener responses', 'accompaniment signals', 'continuers', 'assessments', 'acknowledgments', 'reactive tokens', 'interjectory utterances', and 'recipiency tokens'." (p. 1181). Ward and Tsukahara, however, do not attribute a specific role in conversation to backchannel feedback. Their working definition is that back-channel feedback "responds directly to the content of an utterance of the other, […] is optional, and […] does not require acknowledgement by the other. Note that this definition focuses, not on how these utterances fit into the structure of the discourse, nor on how they are evoked or perceived by the other, but instead on the perspective of the person producing them." (p. 1182). Ward and Tsukahara demonstrate here the context-free, view on short turns in conversation by attributing specific meanings to them, which is different from the CA approach that looks for the meaning of a turn in the co-participants response to that turn: "Back-channel feedback often expresses attention, understanding, or agreement. […] Not all signal attention; some signal boredom. Not all signal agreement; some signal scepticism. Not all signal understanding, often because there is nothing

to understand." (p. 1183). Borrowed from (Yngve, 1970), their interest is in searching for prosodic cues signalling to "a listener that 'it's now appropriate to respond with back-channel feedback'." (Ward and Tsukahara, 2000, p. 1178). It seems that in English and Japanese, "low pitch regions" are such a cue (p. 1178).

Cues that are employed by the prior speaker to invite such short utterances have been examined by Gravano and Hirschberg (2009). Their data was drawn from the Columbia Games Corpus that is a large corpus of task-oriented dialogues. (p. 1019). Backchanneling, according to their definition, is "the production of short expressions such as *uh-huh* or *mm-hm* uttered by listeners to convey that they are paying attention and to encourage speakers to continue" (Gravano and Hirschberg, 2009, p. 1019). Such "backchannel inviting cues" included prosodic and acoustic events, as well as lexical events like part-of-speech. Prosodic-acoustic events were pitch slope, pitch, intensity, inter pausal unit duration and noise-to-harmonics ratio. Their results show on the one hand that the more backchannel inviting cues are used by one speaker, the more likely a backchannel is followed from the other speaker. (Jitter and shimmer, although related to perceptual voice quality, did not affect the probability of back-channel responses.) Regarding the pitch cues, it was found that high-rise contours (H-H%) and low-rise contours (L-H%) account for more than 81% of utterances before backchannels.

Continuer

Research on the same objects, but from a different angle (CA), characterises these short utterances "uh huh", "yeah", "okay", etc. as resources that can be employed by participants to perform specific social actions. Instead of assigning a certain meaning to them a priori (e.g. showing continued interest), CA assigns a social action to them that is grounded in the participant's orientation to them. Schegloff (1982) for example found that an "uh huh" can allow the prior speaker continue speaking and therefore called it "continuer".

Another example is the "mm" token, analysed by Gardner (2001) which can perform different actions and can be called accordingly a "continuer", "acknowledgment" or "assessment" token. They were grouped under the concept of "response tokens" by Gardner (1997, 2001). He follows a similar approach to Gravano and Hirschberg, however with CA as the main method. Gardner demonstrated that these tokens differ in their situational use in conversation. Different communicative functions (continuers, acknowledgments, newsmarkers, change-of-activity tokens) can be assigned to them according to the action they perform in the sequential organisation of talk. Gardner attributes these functions to the response token "mm" and investigates the prosodic properties that are associated with the different categories. The data investigated by Gardner was drawn from a core dataset of Australian-English, supplemented with American-English and British-English datasets. Speakers were engaged in different situations: Talking couples (AUS), people talking at a dinner, people talking while driving in a car (AE). The data also shows speaker preference for specific tokens. For example, there is not a single "uh-huh" in the Australian data (out of 813 tokens). The British data contains 4 (out of 920) and the American 111 (out of 768). The token "yeah" is represented by 60 to 80 %, depending on the dataset (For more detail, see Gardner, 2001, p. 103). This would suggest that the participants in the Australian recordings performed continuer actions other than simply employing "uh huh" (assuming that such actions did occur).

Gardner (1997, 2001) found a dependence between the different conversational functions of the response token "mm" and its prosodic realization. According to Gardner, when produced with a fall-rising F0-contour a "mm" allows the previous speaker to resume talking. If "mm" is produced with a falling F0-contour it acknowledges what the previous speaker was saying. If it is produced with a rise-falling F0-contour it can function as an assessment of the prior speaker's talk (Gardner, 1997, p. 132).

The studies by Gardner and by Gravano and Hirschberg have each addressed interesting questions from different perspectives. Gravano and Hirschberg were interested in the cues in the talk preceding backchannel tokens; Gardner was interested in the cues conveyed by those tokens themselves. It could be argued that Gardner's finding of "fall-rising" F0 contours in continuers

and Gravano and Hirschberg's finding of high rises (H-H%) and low rises (L-H%) before backchannels may have a common origin: the fact that both, backchannels and backchannel invitations share similar pitch contours suggests that prior talk and continuer could be prosodically related.

2.2.2 Prosodic dependency

The notion of prosodic dependency between adjacent turns, hinted at in the previous section, is taken further by Ogden (2006) in a study combining conversation analysis with phonetic observation. Ogden explored the phonetic and interactional resources of agreement and disagreement in assessment sequences. Results show that "'agreement' and 'disagreement' were not found to have unique phonetic properties associated with them." The phonetic resources depended more on whether the second assessment (the second pair part of the assessment sequence) was doing a preferred or a dispreferred action. "In doing a preferred action, the second pair part of the adjacency pair is phonetically 'upgraded' relative to the first pair part. In projecting a dispreferred action, the second pair part is phonetically downgraded relative to the first pair part." (p. 1772). "Upgrading" and "downgrading" phonetically includes here that the F0 is higher or lower, the F0 span is wide or narrow and the speech is louder or quieter than the same properties of the first pair part.

This not only shows that prosody is a relevant resource in the construction of meaning; it also shows that the prosodic pattern of one speaker may be related to the prosodic pattern of the other speaker in immediate adjacency: "A competent speaker matches their own phonetic production to that of another speaker and manipulates the relation between their co-participant's production and their own in ways that have implications for meaning." (p. 1773)

This leads us to another strand of prosodic research in the field of interactional phonetics: prosodic matching

2.2.3 Prosodic matching

The relevance of the sequential analysis of linguistic and phonetic-prosodic features has been demonstrated by Lerner (2002). In environments where choral co-productions of talk are used by participants, matching at a variety of linguistic levels can be observed. Lerner shows "how recipients can make use of the particularities of turn structure, content and context to produce simultaneous, matching TCU elements [talk] in chorus with the current speaker." Most of the matching co-productions are also described by Lerner as being "unison", where for example tempo and pitch of the co-producer matches the "voicing" of the speaker being matched. Different types of action are performed by choral co-productions, such as "conjoined action for a third participant and demonstration of agreement with the current speaker." (Lerner, 2002). These actions are also found in opening and closing conversations and in reminiscing. Most co-productions seem to be used for cooperative and affiliative actions, but Lerner also describes co-productions in other than cooperative and affiliative actions. For example, it is suggested that in turn competition environments "co-production can be a device used for countering the loss of a speaking turn to another participant. Rather than compete openly, one can drop out and take the other's line by co-producing it – and then use that as a basis for continuing one's own line." (Lerner, 2002)

Another case of matching of F0 movements across speakers is described by Walker (2004). Walker (2004, p.119ff) showed that, when granting a first speaker's request, one resource that second speakers use is to match the pitch contour of the request itself, whereas when a request is declined such pitch matching is absent. This was discovered following the initial observation that first pair parts of different action-types like inquiries, requests, assessments, offers and invitations showed a large variability concerning F0 movements. The requests could end in a rise as well as in a fall. These ends of requests made transition to a next speaker relevant. In two examples, the beginning of the request is produced with similar patterns of pitch movements, suggesting prosodic matching (Walker, 2004, pp. 116-122). Walker explains

that "one function of this pitch matching is to display an understanding of the relationship between that talk (the request-response) and the prior talk (the request)" (Walker, 2004, p. 121). It is further suggested that at places where participants have to monitor for transition relevance, they also have to choose the prosodic pattern that they would apply in case they get the speaking floor. This may depend on the action that is planned to be performed and it may also depend on the prosodic pattern that has been produced by the prior speaker.

More work on pitch matching in interaction was done by Couper-Kuhlen (1996). She analysed data from radio phone in programmes. At turn sequences where a speaker at second position repeated literally the words of the immediate prior speaker, not only the words were repeated, but also the prosody was matched. However, the match was further differentiated according to the interactional work it performed and according to the specific prosodic characteristics of that mach. Regarding the prosodic characteristics, the F0 contour could either match on an "absolute" or a "relative" F0 register. An absolute match involved the second speaker to reach the same F0 level as the prior speaker, e.g. 200Hz (irrespective of individual speaker differences). A relative match involved the second speaker to adjust the F0 level to the individual range of the other speaker, i.e. if the prior speaker speaks at his/her mid range, e.g. 200Hz, and the second speaker matches this relatively, this could be achieved by performing his/her mid range, which could be around 100Hz. On the interactional side, Couper-Kuhlen demonstrated that the different types of matches perform different conversational actions: matching relatively, i.e. with respect to the individual's voice range, contextualises verbal repetitions as "quotation", whereas matching absolutely the F0 of the prior speaker contextualises the repetition as "mimicry". In the case of quotation, the lexically repeated words are used as a reference by the person who quotes them. In the case of mimicry, the prior speaker being mimicked becomes a character that appears in the speaker's talk.

In a study that mainly focussed on the interactional properties of continuers was presented by Müller (1996). In Italian conversation the short utterances of a participant, for example "sono d'accordo" (I agree), "si" (yes), "certo" (certainly) or "bene" (fine) were used at locations of continuers and acknowledgment tokens. Müller showed that the actions of these utterances can be differentiated analysing their sequential context and suggested that such short turns do both affiliation and disaffiliation. Furthermore it is suggested that these two actions can be performed by manipulating prosodic features:

> *"Affiliating tokens respond more specifically to important details and to salient prosodic features in the talk they acknowledge. They are more 'matched' responses, hearably more in touch, 'in tune' and 'in rhythm' with the emerging talk of their environment than are their disaffiliating counterparts." (Müller, 1996, p. 163).*

Although the prosodic analysis is less sophisticated than the one by Couper-Kuhlen (1996), it could be shown that short utterances at the position and the lexical makeup of continuers can perform affiliating or disaffiliating work depending on the prosodic matchedness. It is further evidence of a prosodic relationship between turns.

In the most wide-ranging study of this phenomenon to date, Szczepek Reed (2006) demonstrates that speakers routinely orient to the prosodic features used by previous talkers. Types of orientation include prosodic matching (matching of pitch contours, of pitch step-ups, of pitch register, of loudness, of speech rate, of voice quality, of phonetic and sound production), prosodic non-matching, and prosodic complementation. According to Szczepek Reed such orientations can occur in many different types of response, including confirmations, answers to questions, telephone openings and closings, acknowledging next turns, assessments-as-seconds, where an assessment comes at second position, after another first component of a turn (which is not assessing the prior speaker's turn), news receipts: "oh" and related exclamations and disagreements. More recently, Szczepek Reed has proposed that prosodic orientation is central to the sequential management of talk:

> *"Thus, prosodic orientation is shown to be a practice for designing a turn as sequentially continuous, while absence of prosodic orientation may co-occur with sequential*

> *discontinuity in an otherwise potentially continuous environment." (Szczepek Reed, 2009, p. 1243).*

Similarly, in an analysis of parent-child interaction, Wells (2010) concludes that where the child matches the pitch contour of the previous adult turn, this aligns the child with the course of action in progress; alternatively, where the child's pitch contour is noticeably different from that of the preceding adult turn, this initiates a new course of action by the child (Wells B. , 2010, p. 261).

An experimental study by Nilsenová, Swerts, Houteoen and Dittrich (2009) shows that copying behaviour of intonational patterns can be found for example in children, while interacting during a computer game using speech. In this controlled study, the computer game consisted of selecting cards by naming the pictures displayed on them, e.g.: "`I take the elephant`". If the computer voice raised its F0 at the end of the utterance, the child was likely to make a rise at the end of its utterance, too, like in "`I take the crocodile`". A fall by the computer was followed by a fall by the child. This makes it hard to justify that for example requests have to be performed with a rising final intonation, or a falling one (Nilsenová, Swerts, Houteoen, & Dittrich, 2009). The intonational pattern did not follow intonational rules, but was depending on the intonation the child just heard a moment ago.

According to Szczepek Reed, "the majority of research on prosody in conversation to date has focused on exploring the role of individual prosodic features, such as certain types of pitch accent, pitch register or voice quality, for the accomplishment of specified social actions." (Szczepek Reed, 2012a, p. 13) As we have seen in the review of such literature above (Couper-Kuhlen, 1996; Müller, 1996; Walker, 2004; Ogden, 2006; Kaimaki, 2010, 2011) systematic use of prosodic features, but also large variation in the use of prosodic features has been found in participants' implementation of specific social actions. Szczepek Reed argues for an analysis of the participants' "collaborative use of prosody" across turns and across participants. Szczepek Reed demonstrates that "prosody is also relevant beyond the individual sequential location, and instead plays a significant role every time participants place their turns in relation to prior turns." (p. 14) While Couper-Kuhlen (2004) showed that continuation is the default choice, and the beginning of something new a marked choice, Local (1992) found that continuing a previously aborted turn makes the speaker shift back to the previously employed prosodic settings, while a restarting involves a shift to higher pitch than in the prior talk. Szczepek Reed (2012a) further showed that the undertaking of a contextual analysis of prosody can be a fruitful further line of enquiry with reference to "one of the most basic interactional decisions: whether to continue a previously established action trajectory, or whether to start a new one." (p. 13)

2.2.4 Acoustics of prosody

The notions of prosodic matching and non-matching in relation to alignment and non-alignment have been introduced and the boundaries of the units that were used for the following studies are defined. In this section, the parameters are discussed that will be used for the acoustic study. This study involves a technique which can objectively determine the extent to which two speakers' turns match each other in terms of their prosody. This in turn requires a set of prosodic features which can be analysed. Additionally, it is required that these features can be extracted and processed in an automatic way.

2.2.4.1 Acoustic features

There is a wide range of studies which tried to identify acoustic features that are used in studies investigating prosodic cues. But the selection of such features is partly dependent on the analysis technique employed, and partly dependent on the theory adopted in the study concerned. In this section, the various acoustic-phonetic features are reviewed and evaluated as candidate constituents of prosodic matching.

Prosodic parameters

Some authors rely on parameters that have been suggested by the intonation literature, because they seem to correlate with "stress" or "accent". However, the value of such concepts for the analysis of naturally occurring conversational speech is debatable. Nevertheless, most of these parameters have also been used in interactional phonetic research successfully. Therefore a short review follows. It is evaluated how far the parameters that have been suggested by literature working with "clean" speech can also be used for analysing "noisy" speech, i.e. conversational talk.

F0 features and duration were found to be correlates of "stress" (Fry, 1955, 1958 and Bolinger, 1965 cited in Gussenhoven, 2004, Chapter 2), while overall intensity was not found to be correlated with stress (Mol & Uhlenbeck, 1956). Spectral tilt was a further indicator of stressed vowels (with fairly even intensity distribution across the frequency spectrum) and unstressed vowels (downward slope towards the higher end of the spectrum) (Sluijter, 1995). Other influences on stressed and unstressed syllables can be attributed to the vowel quality (more schwa-like in unstressed syllables) and to the duration. "Consonants and vowels in stressed syllables tend to be longer than those in unstressed syllables." (Gussenhoven, 2004, p. 15).

Further factors which influence the selection of prosodic parameters are the recording environment (laboratory vs. conversational setting) and a pragmatic factor: for an automatic analysis as is envisaged in this thesis, those features need to be used, which can be extracted from the acoustic signal.

A complication that every study faces that looks at the fundamental frequency is the mapping of raw F0 measures onto a perceptual scale. One phenomenon that has to be addressed is the non-linear characteristic of that mapping. Another is called "intrinsic pitch" (Ewan, 1975) and relates to variations in F0 that the speaker cannot control for. They are inherently connected to the production of specific speech sounds.

It has long been believed that vowel sounds are mainly determined by the resonance frequencies of the vocal tract. However this is not true. The absolute fundamental frequency of vowels can remain unaltered, as for [i] [a] and [u], while the perceived pitch between them is different. In other words, if we perceive the same pitch, the fundamental frequency of high vowels like [i] is generally higher than for low vowels like [a] and back vowels like [u].

Intrinsic pitch is not limited to vowel sounds, but can also be observed around plosives: "F0 perturbations lead to higher F0 immediately before and after voiceless obstruents than before and after voiced obstruents" (Gussenhoven, 2004, p. 74).

It means that one has to be cautious in the analysis of extracted fundamental frequency contours and one has to be cautious in interpreting the results, if the segmental structure of the utterances is not taken into account. In relation to prosodic matching, pitch or F0 is the most commonly measured parameter in such studies. But there are more than only "intonation contour, pitch register, pitch step-ups" that are all parameters related to pitch. Szczepek Reed (2006) observed prosodic matching in relation to additional features such as "loudness, speech rate, voice quality and sound production." (Szczepek Reed, 2006, p. 35). Szczepek Reed showed that conversational participants do orient towards these features. In her study, the features were obtained impressionistically and for F0 by subjective interpretation of F0 traces. For the current study however, it is envisaged to work with data that can be extracted automatically from the acoustic signals. It is therefore necessary to evaluate how far these features are operationally useful.

The relationship between F0 and intensity

Traditionally, the F0 and intensity parameters have been displayed on separate tracks, as by Couper-Kuhlen (1996), and had to be combined by the observer in an additional cognitive step.

Also in intonation research which is purely based on the F0 contour, it is in exceptional cases allowed or even recommended to consult the intensity contour, as discussed in the ToBI annotation conventions (Beckman and Hirschberg, 2011). For example, it can be unclear what the highest F0 value in a certain syllable is. It may be effected by intrinsic phenomena, e.g. of a voiceless consonant. This would resemble a poor estimate for the analysis of phenomena related to the perceived pitch. If many of these estimates are chosen to evaluate a speaker's pitch range, this can cause errors. Because the voiceless consonants are usually performed with low similarity, Beckman and Hirschberg (2011) suggest that the syllable's amplitude contour can be used to "pinpoint HiF0 within the candidate region." (Beckman and Hirschberg, 2011).

Similar dependencies between F0 and loudness are reported by Kochanski, Grabe, Coleman and Rosner (2005) who assume that "using weights that increase with loudness will emphasize regions that may be more perceptually important." (p. 1043). From a similar standpoint of perceptual salience, for example, Harris (1947) reports that the ability of listeners to discriminate pitch is a function of loudness under certain masking conditions. Hermes (1998b) also used a weighting factor "to assure approximately that speech segments with a higher sound level contribute more to the physical measure [of dissimilarity] than segments with a lower sound level." (p. 75). The weighting factor used by Hermes is the maximum amplitude of the subharmonic sum spectrum. A similar, but simpler weighting factor for pitch similarity is chosen by Rilliard, Allauzen and de Mareüil (2011). They use the signal power instead.

Psychoacoustic evidence for this kind of "trade-off" comes from experiments by Neuhoff and McBeath (1996), who compared in auditory experiments static frequency with dynamic intensity and static intensity with dynamic frequency. They found that a change in pitch can be perceived when a change in intensity occurs and frequency is held constant. In particular, rising intensity can give a percept of rising pitch and falling intensity can give a percept of falling pitch. They further point out that the dimensions of pitch and loudness have been shown to be integral. Stimuli consisting of integral dimensions are "initially perceived as dimensionless, unanalyzable, holistic 'blobs'" (Neuhoff and McBeath, 1996, p. 982). For these reasons, in the present context, it was deemed desirable to use the information of F0 contours in combination with the intensity contours rather than using both as separate and independent parameters.

The combination of F0 and intensity is also employed in interactional phonetic research. For example Walker (2012) analyses F0 traces which have been manipulated by using the information of intensity. Walker "intended to give a visual indication of the perceptual salience of parts of the F0 trace." In the visual representation of the F0 curve, a continuous grey scale was applied, where the F0 contour produced at an intensity of 90 dB was black and of 50 dB was white (p. 144).

The way that intensity is handled in this thesis will be explained in detail below (Section 0).

2.2.4.2 Sequential feature comparison – prosodic matching approaches

In this section, approaches to prosodic matching in the literature are presented with a view to determine their usefulness for the present investigation.

Müller (1996) relied on impressionistic record making using perceptual transcription. For some purposes this is appropriate (see Kelly and Local, 1991), even though it may be inconvenient for the analysis of huge amounts of data. It also raises the question of subjective bias of the annotators. It suggests that a direct comparison of the prosodic contours is a possibility to make the analysis more comprehensive to the reader, as the analyst's explanations can be followed on figures, as it was done for example by Szczepek Reed, 2006 and Couper-Kuhlen, 1996. However, this is not practical with a large amount of data. The ToBI labels of intonational events are another possibility that studies the F0 contour (Silverman, Beckman, Pitrelli, Ostendorf, Wightman, Price, Pierrehumbert and Hirschberg, 1992). However, the ToBI system suffers from low inter-rater reliability, as Wightman (2002) discussed and studies of Grice, Reyelt, Benzmüller, Mayer and Batliner (1996), Gut and Bayerl (2004) and Pitrelli, Beckman and Hirschberg (1994) demonstrate. In order to increase objectivity and the amount of data that

can be investigated, it is important to look for other possibilities of evaluating the similarity of prosodic features than impressionistic or non-impressionistic annotations.

Automatic acoustic analyses could be such a possibility. Kousidis, Dorran, Wang, Vaughan, Cullen, Campbell, McDonnell and Coyle (2008) conducted a study to evaluate the convergence of speakers in unconstrained dialogue. They used a technique which they call "time aligned moving average" (TAMA) on 24 acoustic parameters. Time frames over which the average was taken are 10, 20, 30 or 60 seconds long and had 50% overlap. This automatic technique shows one possibility of processing large amounts of data in an objective way. However, it can be questioned if the approach can caption the qualitative aspects of conversational actions. For example, the authors consciously avoided a turn-by-turn analysis, as the frames may contain different types of utterances simultaneously. The first part of a frame might for example contain a declarative, while the other part of a frame might contain an interrogative or back-channel. As Kousidis et al. (2008) state that back-channels have "inherently lower pitch" and "questions having higher pitch than statements" (p. 1693), the differences get averaged out and are therefore lost. This is one of the reasons they claim why pitch range does not converge and average pitch only slightly converges, while intensity and speech rate do indeed converge. They show one example with frames of 10 seconds duration. The speakers occasionally diverge one from the other in terms of intensity, for example when one of the speakers poses "a sudden question, characterized by higher intensity, [or with productions of] non-lexical very short utterances [, such as] ("wow")." (p. 1694). The aim of this thesis is to capture exactly these local phenomena and should therefore avoid the averaging of prosodic parameters over longer time stretches.

In a related study by Kousidis, Dorran, McDonnell and Coyle (2009) it is stated that due to this "complex structure of spontaneous speech, the most recent utterance may not be the most relevant." (Kousidis et al., 2009, result section). This is speculative: as the previous sections have shown, every utterance, and especially the most recent one, is relevant for the unfolding of the conversation and for its participants. It is quite possible that the prosody in this utterance is also of importance. Although the approaches of Kousidis et al. (2008, 2009) have interesting aspects as for the choice of prosodic features, "prosodic convergence" is a different concept from prosodic matching. Although prosodic convergence cannot be used for evaluating how participants match each other on a turn-by-turn basis, it should be kept in mind that participants tend to match the overall prosodic properties over longer stretches of talk, such as 10s up to 1 minute.

Other research has used shorter stretches of the duration of a single utterance. Using non-conversational speech, Hermes (1998b) provided measures of physical (dis)similarity of intonation contours. One purpose of Hermes' study was to find a measure of prosodic similarity to teach intonation to persons suffering from deafness. The production of an utterance from the participant needed to be compared with a template utterance.

The underlying material, taken from Hermes (1998a), was "spoken sentences chosen in such a way that a relatively wide variety of pitch contours was realized" (p. 65). These contours were stylised using pitch-synchronous-overlap-and-add (PSOLA) (Hamon, Mouline, & Charmentier, 1989) and resynthesised, thereby removing all voiceless segments. They all had identical timing properties (the same duration). If this technique were to be applied to conversational speech, the requirement of equal durations of the two compared intonation contours would be a serious disadvantage, because the turns from speakers in real conversations rarely have the same duration.

In order to overcome this limitation, Rilliard, Allauzen and de Mareüil (2011) used the dynamic time warping technique, which allows comparing the similarity of contours with different overall duration. Rilliard et al. also modified Hermes' correlation coefficient by using the signal power (instead of the subharmonic-summation (SHS), as it was used by Hermes) for the weighting factor across both speakers.

The measures by Hermes (1998a) and Rilliard et al. (2011) capture similarities in F0 contours well, using a weighting factor for the similarities or differences in the F0 contours. However, the correlation technique (the Hermes measure) and the dynamic time warping technique cannot

cope with areas where no or only sparse readings of the F0 are available. This may be the reason why Hermes and Rilliard et al. used linear interpolation techniques to obtain continuous F0 and intensity signals.

Because voicelessness is a major issue with data from spontaneous conversations with naturally occurring talk, it has strong implications for the prosodic similarity metrics described below. It is highly desirable to avoid interpolation for missing F0 values which can introduce an uncontrollable source of errors.

It is notable that all the studies mentioned in this section used normalisation techniques for the F0 parameters. The signals from the individual speakers can therefore be compared on equal terms. The interactional analysis by Couper-Kuhlen (1996) also found a difference in the orientation of participants to absolute vs. relative pitch matching, indicating that normalisation is appropriate to account for the "relative" matches. However, the "absolute" matches are hard to model after normalisation has been applied. This would then result in a non-match on the relative scale.

2.2.5 Gestural resources

We might not always be aware of gestures while we talk, but they occur almost all the time. We move our head, arms, hands, eyes, eye-brows, and other parts of the body while being engaged in conversation. Instances where the movements are true gestures and the way how gestures are to be classified are dealt with below. One main issue for researchers has been the alignment of gestures with speech events. Another issue is whether gestures are employed to support speech in its production, or whether gestures are also relevant for the recipient. Most of the psycholinguistic literature about gesture focuses on those gestures which are performed with hands and arms to depict a complex scene, such as McNeill's (1992) example where a subject reiterates a scene from a cartoon: "`he grabs a big oak and bends it way back`". A grasping motion and a pulling motion back down accompany the words "`bends it`". As interactionally-relevant movements of hands, the head (nods, shrugs, etc.) and facial displays (frown, raised eyebrows, etc.) are the main focus of the gestural part of this study, a connection of CA with head gestures is established where possible, with reference to the relevant literature.

2.2.5.1 Speech and gesture production

This section gives an overview on the use of bodily movements and their synchronisation with speech events. There is evidence from psycholinguistic experiments that gestures facilitate speech production. For example, Habets, Kita, Shao, Özyürek & Hagoort (2011) investigated how the semantic integration of gesture and speech works. By modifying the speech onset after the gesture and by introducing semantic congruent vs. incongruent stimuli, they sought to answer the question in what time frame gesture and speech information are integrated best. Their findings using EEG recordings suggest that after a time lag of 360 ms before the speech onset, the interpretation of the gesture is fixed.

Another psycholinguistic study (So, Kita, & Goldin-Meadow, 2009) looks at the link between gesture and speech and supports the view that gestural representations are shaped by the online linguistic choice of the speaker, rather than by pre-determined language-specific schemas. An experiment where subjects were required to re-tell stories and specify referents showed that whenever speech failed to uniquely specify a referent, gesture should also fail to specify the referent. So et al. "suggest that speakers did not use gesture to compensate for speech because gesture and speech are part of a single, integrated system." (p.121). In contrast to what one might have expected, that speakers would use gesture to convey information that is not found in their speech, speakers did not use gesture to compensate for the referents they failed to specify in speech. This suggests that "gesture is used redundantly with speech." (p.122)

The two studies above indicate that gesturing and speaking are tightly linked and that speech and gesture production processes share a common computational stage: "gestures are

informationally and temporally well-coordinated with the concurrent speech." (Kita, Özyürek, Allan, Brown, Furman & Ishizuka, 2007, p.1213)

Kita et al. (2007) argue against the Free Imagery Hypothesis according to which "iconic gestures are generated from imagery that is formed 'prelinguistically', that is, before linguistic formulation processes" (p. 1214) and argue for the Interface Hypothesis, "according to which gestural representation is shaped in the process of organising information for speaking." (p. 1216). The aim of the study was to test whether gestural representation is shaped by the online linguistic choice (Interface Hypothesis) or by pre-determined language-specific schemas (Habitual Conceptualisation Hypothesis). The experimental setup was chosen to elicit from subjects speech and gestures that indicate the manner and the path of objects. Thereby, the choice was open to describe them in so-called tight clauses (both manner and path in one) or in two separate clauses; the same on the gestural side (both manner and path in one combined gesture) or in two separate gestures. Results showed that tight clauses were likely to lead to combined gestures, while separate clauses were likely to lead to single-information gestures. This suggests that speech influences gesture. The syntactic link of manner and path was also reflected in the gestural link of manner and path. If manner and path were not linked syntactically, they were also separated in gesture. Kita et al. conclude "that the speaker's online choice of the clause type had a unique and independent influence on gestural packaging of Manner and Path. This provides support for the Interface Hypothesis, and they are at odds with the Habitual Conceptualisation Hypothesis." (p.1231). It is argued that the theoretical implications for speech production are that the "gesture's online sensitivity to syntactic packaging of information suggests that conceptual message representations and syntactic representations are generated interactively during speaking." (p.1231).

However, not all gestures are performed in synchrony with speech. For example Efron (1941) differentiated between gestures which derive their meaning in conjunction with the verbal utterance and gestures which derive their meaning independently from verbal contents and can be performed in isolation. The former are for example "batons" which indicate the rhythmic structure of an utterance. The latter are illustrative and symbolic gestures, such as the "OK" sign (index finger and thumb build a circle, while the other fingers spread apart).

A classification scheme for gestures, introduced by Kendon (1982), was arranged on a scale by McNeill (1992) from "gesticulation" over "emblems" and "pantomime" to "sign language". According to McNeill, gesticulation is obligatorily accompanied by speech, while emblems are only optionally accompanied by speech. For example the emblemic "OK" gesture can also be accompanied by a verbal "okay", but does not have to. The other two (pantomime and sign language) are not the focus of this study. Here, the gestures that develop spontaneously over time are analysed, may they be accompanied by speech or not.

There is a slight confusion with Kendon's definition of "gesticulation" and its wide synonymous use as "gesture", e.g. by Loehr (2004): "Definitions for all these types of movements were provided by Kendon (1982), who prefers the term *gesticulation* for what [Loehr] (in keeping with most of the authors) will refer to as *gesture*." (Loehr, 2004, p. 7). There is the observation of body movements, such as nods which are also termed "gestures". Head nods can either be accompanied by speech, as we will see in our data collection or they can occur alone (Stivers, 2008). However, they are supposed to be used for the same purpose, i.e. to manage the turn taking between speakers. Ekman and Friesen's category including nods, eye contacts and postural shifts are the "regulators" to "maintain and regulate the back-and-forth nature of speaking and listening between two or more interactants" (Ekman & Friesen, 1969, p. 82).

Gestures are similarly important on the recipient side for language comprehension (Holle, Obermeier, Schmidt-Kassow, Friederici, Ward & Gunter, 2012). In an experiment on ambiguous syntactic structures, beat gestures helped in the disambiguation process. Holle et al. suggest that "unlike a sentence context or a visual scene, beat gestures do not operate on a semantic level. Instead, these hand movements can emphasize a certain phrase irrespective of their concrete form. It is just important at what time the movement occurred, not what it looked like." This suggests that gestures are helpful for the recipients to evaluate the important parts that have been emphasised by the main speaker. Holle et al. argue further that "gestures are

always part of the communicative exchange, and may therefore serve as a very natural and powerful cue to shape the interpretation of spoken utterances." (p. 9) This suggests that gestures are important for both the production side and the recipient side.

Not merely the detailed characteristics of individual gestures have been analysed, but also their contextual use. For example, the analysis of head and torso movements leads Scheflen (1964, 1968) to the observation of the use of gestures in the proximity of other gestures. A posture shift of one interactant was followed by the same gesture of the others. Wundt (1973) goes one step further, as Kendon (2004) has put it: "One person's imitative responses to another's expressive movements become *answering* movements, not just imitations." (Kendon, 2004, p. 58). This kind of alignment of gestures from one participant to the gestures from another participant and the relationship of individual gestures in face-to-face interaction is discussed below.

2.2.5.2 Specific gestures – specific actions

The search for regularities in gestures in relation to social actions is demonstrated by researchers of other interests than Conversation Analysis, such as Adam Kendon and David McNeill, who analysed gesticulation mainly from the perspective of the speaker, e.g. a narrator of a story. According to Kendon (2004), "interactive or interpersonal functions of gestures [...] that regulate turns at talk [... are] important and interesting", but a "systematic discussion seems lacking" (p. 159). The analysis of adjacent turns therefore requires an approach to gestures which takes into account the actions that are done by and to interlocutors.

One such study that investigates how specific gestures are used in order to perform specific social actions was presented by Schegloff (1987). A distinction between horizontal and vertical head movements is made: "a horizontal or lateral head shake can have at least three distinct uses: as a marker or expression of the negative, of disagreement, and/or of intensification." (p. 106). "The vertical shake or nod has a major use as a "continuer" or indicator that a recipient of speech understands that an extended unit of talk is in progress and should continue." Schegloff gives such gestures the status of a turn, as "an ongoing speaker may leave a bit of a silence into which such a continuer may be inserted, thus making the nodder into a virtual speaker at that moment." (p. 106). That is also one reason for adding the gestures in separate lines in the extracts above (Section 2.1.3), as they can be considered as turns in their own right.

A similar perspective on gestures from the CA point of view is provided by Stivers (2008). Looking at storytelling, Stivers relates the nodding gesture of one speaker to the surrounding talk of the other speaker and assigns a specific type of social action to it; affiliation or disaffiliation. This results from an analysis of nods according to their direct interactional context. The action that the nod accomplishes is evaluated from the interlocutor's reaction to it. Nods at mid-telling position are used to "help convey that tellings are on their way to preferred affiliative uptake at story completion." (p.52) As the nod projects that the story will receive affiliative uptake at its completion, the nod is already treated as if it was affiliative at the time it was produced, i.e. at mid-telling position. However, if the nod occurs at the wrong position, namely at the end of a story, where for example a positive assessment would be a preferred uptake, it is treated by the story-teller as ill-fitted. Thus it is treated as if it was disaffiliative. Stivers found that therefore, "at ends of tellings, story recipients and tellers alike do substantial work to *avoid* nodding as a final response." (p. 49)

While Stivers used a simple binary classification of nods, indicating if a nod was present or not, Whitehead (2011) analysed nods in more detail and found differences in social actions which can be attributed to differences in the details of the conduct of nods. Examining the use of speaker's nods in "minimal post expansions" three distinct types of nods were found. A nod which is used alone or in conjunction with a "change-of-state" token (e.g. "oh") is shown to treat the prior utterance as news, in agreement with the analysis of the verbal change-of-state token. Such a nod is designed as a long and expansive up and down movement. A different type of nod ("acknowledgment nod"), which marks the receipt of an utterance without treating it as news, is designed as "markedly less expansive in terms of their amplitude and total duration" compared to the first type of nod and consists of several up and down movements rather than

just one expansive one (Whitehead, 2011, p. 113). The third type of nod has features of the previous two types and seems to acknowledge "dispreferred" news, however Whitehead makes no specific details of the realisation of this type of nod available. Nevertheless, Whitehead's detailed analysis of head nods shows that distinct types of head nods can be used to perform specific actions in turn sequences.

McClave (2000) performed microanalysis on conversations between native speakers of American English. Many linguistic functions of head movements were identified, including affirmation and negation, inclusivity and intensification and marking of switches between direct and indirect speech. The most relevant finding for the current thesis is however, that head nods were used by participants to receive feedback from their co-participants. In the analysed data, "the listeners interpret speaker nods as requests for input". And such input can either be verbal backchannels or non-verbal backchannels, such as nods. McClave showed that head nods "clearly function interactively." (p. 876). Additionally it suggests that nods can occur in sequences of nods, where a nod from one speaker may be conditioned by a previous nod from another speaker. This may even have consequences for the interaction.

2.2.5.3 Gestures in adjacent turns

The studies mentioned above focused mainly on gestures involving the head, especially nods. In this section, the scope of gestures includes movements of other parts of the body, such as hand gestures. Following McClaves' (2000) finding that specific gestures can be followed by the same gestures of the next speaker, studies that analyse gestures in a sequential setting and analysing gesture sequences are presented here.

Holler and Wilkin (2011b) summarise work on gestures functioning in three ways: gestures constituting addressee feedback, gestures for eliciting addressee feedback and gestures in response to addressee feedback. Recent gesture research has established some knowledge about detailed use of the first two types of gestures and suggested their communicative function according to what the gestures were used to do or to explain. But, as Holler and Wilkin (2011b) argue, "we still know comparatively little about the role of gestures in the actual *process* of communication." (p. 3522). What Holler and Wilkin call "process" means in the context of gestures how speakers adjust their gestures in response to specific types of addressee feedback. The recording setup involved one speaker who was engaged in narrating a story. The other speaker was asked to give "feedback" on the story. The feedback was scripted and contained the following four types:

1) "Request for further information due to lack of clarity/detail"
2) "Expression of non-understanding – provision of incorrect interpretation"
3) "Expression of non-understanding – request for speaker to repeat or clarify what she said"
4) "Seeking confirmation that understanding is correct – provision of correct interpretation". (p. 3526)

Questions Holler and Wilkin sought to answer were how often the prior speaker repeated the previous gesture or how often the next gesture was modified in response to feedback. And with regard to a modification, the question was if this modification was making the gesture more "communicative" after the feedback (p. 3524). In the first three conditions, when repair or elaboration was initiated, they found that "speakers continued to use gesture to the same extent before and after feedback" (p. 3532). In the fourth condition, where confirmation of a correct interpretation was initiated, this confirmation was mainly given orally. No repetition of the gesture seemed to be required. The experimental setting limited the interaction to one-sided storytelling, and therefore limits the conclusions to that narrow scope. Additionally, the gesture properties were measured quantitatively including gesture precision, size and space, rather than their type of gesture. Nevertheless, the approach taken by Holler and Wilkin indicates that gestures are dependent on the surrounding conditions. It also indicates that the sequential analysis of gestures may help to identify actions which are performed by other speakers than the

gesturer him/herself. Although Holler and Wilkin do not mention this, it might be possible to tell from the behaviour of the gesturer, before and after the addressee's feedback, which of the four actions were performed by the feedback. From their study, it seems that the type of feedback has an influence on the gesturer whether to continue or discontinue the gesture prior to the feedback.

Two studies looking from a qualitative standpoint of analysis at the use of gestures used by the participant of a conversation who is currently not the turn holder, are (Lerner, 2002) and (Schegloff, 1984).

According to Lerner (2002), Schegloff (1984) distinguishes three sequential environments where the "nonspeaker" uses hand gesticulation:

> *"1. gesturing by a nonspeaker as a way to make a move for a speaking turn, 2. gesturing in lieu of talk as a way to communicate without interrupting a current speaker and 3. maintaining a gesturing pose after yielding to an interrupting speaker to show that one considers their own turn to still be in progress." (Lerner, 2002).*

The third point demonstrates that a first speaker can show continuation on the own course of action by continuing the gesture after a turn from another speaker. In such a case the other speaker's turn was treated as an "interruption".

> *"Such at-that-moment nonspeakers may hold a gesture that was in progress at the point of interruption to show that they consider their turn still in progress and intend to resume after the interruption." (Schegloff, 1984, p. 271).*

This suggests that the prior turn of the speaker being interrupted has not finished and requires continuation. Schegloff showed that this can be achieved by a continued gesture. Looking at this finding in the light of verbal continuation as it is allowed by verbal continuers (see Section 2.2.1.2 above), it seems reasonable to argue that even if the turn from the other speaker was not designed as interruption, the speaker who initially occupied the turn continues with the prior gesture. This may be the case, especially if the "interrupting" turn was designed as continuer, i.e. as a turn that allows the prior speaker to continue on the prior course of action. One possibility of a prior speaker to treat it this way is then to continue verbally – or gesturally – the prior verbal talk or gesture.

Lerner (2002) concludes that "co-production, as a feature of the organization of social interaction, does not seem to be limited to talk, but can be accomplished through gestural matching."

2.2.5.4 Mimicry and other potentially related activities

The section above dealt with gestures in adjacent turns and the consequences they can have on the interaction. In this section the notion of gestural matching is added. How are gestures analysed when they are related to the immediate gestural context, i.e. the preceding gesture of the prior speaker? Terms that have been used in this context are mimicry (Kimbara, 2006), (Parril & Kimbara, 2006), convergence (Kimbara, 2008), imitation (Paukner, Suomi, Visalberghi, & Ferrari, 2009), copying, mirroring, shadowing (Tannen, 1989) and mutual monitoring (Goodwin M. H., 1980; Goodwin and Goodwin, 1987).

Some of these terms are used to describe equivalent ideas. Mimicry, copying, mirroring and shadowing describe the phenomenon of close proximity in the time domain. However the term convergence describes a process which develops over a longer stretch of time than an immediate change from one turn to the next turn.

Mc Neill (2008) observed mimicry in hand gesticulation. He states: "Mimicry obviously is a social interactive response. Less obviously it is also a tool for comprehending the other person, which may be one reason it occurs in the first place. Mimicry is a form of materialization, the materialization by a listener of another person's gesture." (McNeill, 2008, p. 8).

Thus, showing one's comprehension of the other person seems to be one purpose for which mimicry is employed in conversation.

Mimicry has also been investigated from an evolutionary perspective. Lakin, Jefferis, Cheng and Chartrand (2003) argue that non-conscious mimicry gives an evolutionary advantage. They suggest that it "serves to foster relationships with others" because "nonconscious behavioral mimicry increases affiliation." (p. 145) Non-conscious mimicry is not necessarily limited to gestural mimicry. It can also be mimicry of accent, speech rate and rhythms (p. 148). However, it is not explained where this affiliative characteristic of mimicking comes from and how it works. No data is reported, which leaves it at the stage of an interesting hypothesis.

Research in mimicry has also inspired studies using a robot as conversational partner that could copy facial gestures (Riek, Paul, & Robinson, 2010). In their experiment, the robot copied the visual behaviour of people who were asked to tell the robot a story. The hypothesis was that the more gestural features are copied, the more positive the participants will rate their interaction with the robot. However, this could not be verified and the null-hypothesis that copying does not have an influence on the perception of the interaction as positive, remains valid. Nevertheless, Riek et al. made an unexpected, but interesting observation of co-nodding: Two participants "nodded, the robot nodded in response, and then the participant nodded to acknowledge the robot's nod." (p. 105). This suggests that gesture copying may still occur, even if it is not valenced as something positive or negative.

An interactional study on storytelling involving real participants in naturalistic conversation was done by Selting (2010). It is shown that in complaint story telling environments, participants either match prior formulations in structure and prosody (p. 241) or they don't. They also accommodate facial expressions, for example "by also enacting raised eyebrows" (p. 244) if the prior speaker did. According to Selting, one can observe sequences "for the collaborative treatment of affectivity in climaxes of complaint stories, with affiliative responses by the recipient." (p. 241). These can manifest in two adjacency pairs. In the first pair, the story teller shows "anger" or "indignation" at the story's climax which are matched by the recipient. In the second pair, the story teller "expands on the climax by evaluating the complainable [...] slightly weaker, to which [the recipient] responds a bit weaker as well, i.e. exactly matching. This shows the interlocutors' precise monitoring and management of their displays of affectivity; they orient to each other and adapt their displays towards each other." (p. 244).

We should note that after the affiliative (matching) display of the recipient, the first speaker continues on her prior course of action, i.e. by expanding and elaborating, as this is relevant for the definition of the action pair alignment and non-alignment.

In a contrasting example, "when the recipient provided responses not matching in affectivity, the storyteller could be seen to continue and even upgrade her own subsequent displays of affective involvement, thus creating further in-situ opportunities for the recipient to respond in a better matching manner." (p. 255) Therefore, "The ***recipients' withholding of affiliative responses*** to the climax of the story with the display of emotive involvement causes the storyteller to ***expand*** the storytelling [...] The expansions may display even stronger and clearer emotive involvement, presumably in order to again and more clearly elicit affiliative responses." (Selting, 2010, p. 271, bold marking in the original)

These affiliative responses are mainly characterised as matching responses in terms of phonetics, prosody and gesture (or facial expression):

> *"Here, where the interlocutors produce matching displays of affect, [the storyteller] treats [the recipient's] displays of his agreement and affiliation as unremarkable and immediately continues her telling." (p. 272). But, "there is* ***no simple "mirroring" of affects*** *in natural social interaction. Logically for interaction, recipients would rather be expected to perform the complementary task of perhaps soothing, calming down, de-escalating the speaker's emotive involvement." (Selting, 2010, p. 272, bold marking in the original)*

Selting's study shows that both the *prosodic matching and non-matching* of the prior speaker's utterance and the *gestural matching and non-matching* of the prior speaker's facial displays can be employed by the second speaker in order to perform specific actions in specific

environments. There, the environment is complaint storytelling, where matching resulted in continuation of the action, while non-matching resulted in expansion and upgrading.

2.2.5.5 Summary

The nonverbal part of a conversation is as important as the verbal part. Applying these findings to the envisaged study of social interactions where one participant allows the continuation on the prior agenda or initiates a change of action, this decision of a first speaker, whether to continue on the prior agenda, may not only depend on the lexico-syntactic and prosodic characteristics of the second turn. It may equally depend on the gestural properties of the second turn or a combination of several of these factors. Three basic approaches for gesture analysis can be established:

1. Kendon (2004) suggests that gestures are used in addition to speech in order to transfer meaning from the speaker to the interlocutors. For example nods are claimed to work as emblems which "can substitute for speech because meaning is recognised in the sign itself" (Orton, 2007). This suggests that the gesture can gain its function by its shape itself and by its co-use with speech.

2. Stivers (2008) found that nods can do affiliating work in the environment of storytelling. But nods do not always have the same function. It also depends on the location where they are placed in the specific context. If a nod is placed at the "wrong" position, the action it performs turns into the opposite and the nod does then disaffiliating work. This suggests that the communicative function of a gesture depends on its location in a sequence, additionally to its shape.

3. The communicative function, or action, may depend on the degree of the relative match of the said gesture with the immediately preceding one (Selting, 2010).

Applying the scheme of the social actions alignment and non-alignment (see Chapter 4), the treatment of the target turn as a turn that allows continuation on the prior agenda (alignment) or not (non-alignment), may directly, or at least partially, depend on the gestural match of the second speaker with the first speaker. In analogy with the acoustic-prosodic comparison of adjacent turns (see Chapter 5), it seems to make sense to look for the gestural matches in these turns. When no preceding gesture exists, the gesture in question can be viewed in relation to that non-existent, or "none" gesture.

2.3 Units

The research on prosody and gesture, as well as prosodic matching and gestural matching, shades into the matter of choosing the units which need to be employed in order to investigate all these phenomena. The conversation analytic way would be to take those units which are employed by the conversational participants, but this is a complicated issue, as is demonstrated by arguments about turn constructional units (TCUs) and transition relevance places (TRPs) in Schegloff (1996) and Selting (2000) and the discussion of units in Szczepek Reed (2010b) and Szczepek Reed and Raymond (forthcoming).

The CA view is that the most basic unit by which talk is organised is the turn constructional unit (TCU), as it was introduced by Sacks et al. (1974). According to Schegloff (1996) it is demarcated by transition relevance places (TRPs), which make transition from one speaker to a next speaker relevant. Whether an uptake by a next speaker occurs or not however depends on various factors. One factor may relate to lexical, grammatical and pragmatic cues, another may be language specific.

The factor prosody is addressed by Selting (2000) when referring to "turn-final pitch [...] signalling possible turn completion". In general it is suggested that some turn-yielding and turn-holding cues are prosodic and a "speaker uses falling or rising final pitch as a possibly turn-yielding pitch. In marked cases, when speakers intend to hold the turn for more than one TCU,

they need to use special turn-holding devices which project more to come." (p. 510). It is then the task of the co-participant to monitor for these cues in order to choose either to start an own turn or to leave the floor to the other speaker.

Another factor for speaker uptake is syntactic, for example the *if-then* construction, which is used to stretch a turn across potential prosodic boundaries. Selting (2000) says about findings by Lerner (1996): "An instance of longer TCUs is also given in what have been termed "compound TCUs." In perfect agreement with the turn-taking model, Lerner 1996 analyzes *if-then* and *when-then* constructions as "compound TCUs" – even if a prosodic break, signaling preliminary component completion, displays the entire construction in two prosodic or intonation units." (Selting 2000, p. 481) The *if* part may be delimited from the *then* part prosodically, but connected syntactically. Selting concludes:

> *"In this view, the TCU is by no means coextensive by definition with linguistic units defined in terms of syntax and prosody. It can be coextensive with single sentences, clauses, phrases, etc.; but it can also be much longer than one such unit. At the same time, according to this view, these "big packages" must contain some other kind of "unit" below the TCU." (Selting 2000, p. 486)*

Selting further argues that "A turn ending in a TRP can thus be built with one TCU or more than one, and TCUs can be built with one or more intonation units." (p. 490)

The term "intonation unit" is however not very precise; but nor is "chunk" (Selting, 2000) or "intonation phrase-like chunk" (Szczepek Reed, 2010b, p. 191). Szczepek Reed argues that "chunking" seems to be oriented to by participants as an "interactional strategy, employed for the structuring of turns" (Szczepek Reed, 2010b, p. 204) and that "an observation of next participants' treatment of chunks as chunks can be relatively straightforward." (p. 200). However, no hard criteria have been proposed for the delimitation of "chunks".

Earlier, Szczepek Reed (2006) has demonstrated that participants orient to prosodic patterns used by co-participants for example by matching and non-matching these patterns. It could be argued that such matching of patterns is implicitly an orientation to underlying units. But it would be important to demonstrate that participants also orient to these chunks interactionally, i.e. that the orientation has specific interactional consequences. It needs to be demonstrated what actions the prosodic matching is doing and to what actions it contrasts. This would then reveal that these units are interactionally relevant or an "interactional strategy". However, as was shown by Selting, the prosodic units and the syntactic units are sometimes at odds with each other and do not consistently indicate TCU boundaries: "prosody and intonation cannot be seen as providing a unique criterion overriding, e.g. syntax. The TCU is not identical with an "intonation unit" or "prosodic unit"." (Selting, 2000, p. 490). Orientation by participants towards each other and the context may reveal the relevance of these units. Participants obviously orient to TRPs, but the questions whether participants also orient to TCUs, or other prosodic sub-units are still open.

The aim of this study is to evaluate the prosodic matching hypothesis. Therefore it is impossible to choose the underlying unit as an entity which is defined on the basis of prosodic matching and non-matching without introducing circularity. Because the debate on intonation units is still ongoing, the unit for the investigation in this thesis can be drawn on other literature that defines intonation units, but still according to accessible rules which delimit its properties, i.e. its boundaries. The "intonation phrase" is such a neutral unit. Its boundaries are defined according to "tone", "tonicity" and "tonality" (Wells J. C., 2006, p. 6 ff.). According to Wells, there are "different kinds of chunking possible". One possibility is to delimit intonation phrases (IPs) using clauses. Another possibility is to split an utterance according to "pieces of information". If an information piece is packaged as a single IP, it carries "one intonation pattern". Several information pieces carry several intonation patterns. This segmentation into information pieces is called "tonality" or "chunking". In contrast to tonality, "tonicity" is used by speakers to "highlight some words as *important*" (Wells, J. C., 2006; emphasis in the original). Such important words are highlighted with a so-called "accent".

The working definition for the intonational phrase (IP) includes phrases that have not more than one accent. This accent may be realised with a pitch movement, increased intensity, or some other parameter putting emphasis on a certain stretch of that phrase. If no such accent can be identified, the IP is delimited by the next grammatical phrase boundary.

The purpose is to select those parts of the utterances from the interacting speakers which can be used for interactional and later acoustic analyses. Utterances or clauses from interacting participants may or may sometimes not carry an accent. Therefore, the definition of "intonation phrase" (Wells J. C., 2006) is slightly adapted to fit the needs of chunking. Additionally, if this chunking was made dependent on the analysis of tones and accents, this could potentially introduce circularity into the interactional and the acoustic analysis later. The thesis should provide possible answers to the work of prosodic resources on turn taking without implying too many assumptions.

2.4 Proposed methodology – Combination of CA and phonetic-prosodic-gestural analysis

Conversation analysis includes an inventory of procedures and methods for investigating qualitatively the actions performed by participants in natural conversations. Although these methods are objective from the CA point of view, there are other research disciplines which are more critical about CA findings because they rely principally on the collection of impressionistic observations based on orthographic transcripts. In interactional phonetics research, objective instrumental measurements of acoustic properties have been combined with CA techniques, as for example by Ogden (2006, 2010), Walker (2004) and Local & Walker (2005). The fundamentals of this approach are explained by Local and Walker (2005) and its relationship to speech technology research is described by Kurtić, Brown and Wells (in press). However, further analysis methods could be adopted from laboratory phonetics, to improve the validity and reliability of the acoustic analysis and to automatise analytical procedures in order to investigate large corpora. Even though such laboratory methods have been designed for the analysis of speech in clean acoustic conditions and the adaptation to naturally occurring talk is challenging, if successful, this would increase the robustness of phonetic analyses of talk-in-interaction.

With regard to the analysis of social actions, which are to be correlated with auditory (prosodic) features, we propose to build on the concept, shared by CA and interactional phonetics, of conversation being organised sequentially turn by turn. Alongside this sequential analysis of social actions, a sequential analysis of prosodic features is proposed. The social actions first have to be established on interactional grounds by building data collections. These represent the interactional categories which can be evaluated on acoustic and gestural grounds. The prosodic analysis of the acoustic signals extracts features (F0 and intensity) and compares them sequentially across speakers and turns. A further aim is to perform a similar analysis of the gestural modality by using video signals or signals from motion tracking devices. The main results to be presented here were achieved on the basis of gesture annotations which have been done by hand / sight. Preliminary investigations of the data obtained by the motion tracking prototype are promising and will be presented in Section 7.3.

2.5 Discussion - implications for the present study

Previous research on intonation suggests that there is a connection between the choice of pitch contour, span, register, tempo and rhythm and communicative meaning in general. For example, it is claimed that specific contours have specific meanings similar to an intonation lexicon (Levelt, 1989). This might be true for the environment of speech artificially designed in the laboratory and exposed to subjects in experimental conditions. It is suggested that findings

based on these experiments are hardly true in real conversations (Geluykens, 1987) and can be challenged by qualitative analysis of naturally occurring talk. In return interactional research showed evidence of the understanding that the choice of pitch contour is managed by interacting speakers locally, depending on the choice of pitch contour of the immediately preceding speaker (Walker (2004), Ogden (2006), Müller (1996), Couper-Kuhlen (1996), Szczepek Reed (2006, 2009), Wells (2010), Nilsenová et al. (2009)). To test this hypothesis more robustly, a prosodic analysis which is purely based on the acoustic signal is needed.

With regard to short responses such as "uh huh", previous work has done reasonably well in identifying acoustic and visual cues that are inviting "backchannels" on the one hand (Gravano and Hirschberg, 2009) and in identifying such properties for backchannels / response tokens themselves on the other hand (Gardner, 2001). Gravano (2009) investigated the features of the backchannels regarding the lexical, discourse, timing, acoustic and phonetic domains. However, the comparison is missing on an individual basis of each response token with the immediately preceding talk of the other speaker – in the auditory / acoustic and in the visual / gestural domain. An investigation is missing about the cues which follow or precede other cues, being similar or different. Regarding Gravano and Hirschberg's study, it would have been interesting to see if those backchannels which follow other F0 cues than rising pitch, e.g. L-L% (low-falling) contours had sometimes the same organisation of features (i.e. also low-falling).

Instead of a two step analysis: first investigating cues inviting backchannels and second investigating acoustic-linguistic properties of backchannels, the acoustic analysis described in this thesis will be a one-step approach. This is a direct comparison of the acoustic and gestural properties of the second speaker with the immediately preceding acoustic and gestural properties of the prior speaker. This is in line with one of the conclusions of the results of a perceptual study comparing pure tokens of "okay" and tokens embedded in context of two turns, including the token "okay" described by Gravano, Benus, Chávez, Hirschberg and Wilcox (2007, p. 806 f.). Subjects oriented towards different cues depending on whether the context was given or not: "... these results suggest that contextual features might override the effect of most acoustic, prosodic and phonetic features of okay". Another important conclusion driving their future work is to investigate: "how subjects adapt their choice and production of cue phrases [another perceptual category in their study] to their conversational partner's" (Gravano et al., 2007, p. 807). This also suggests that matching vs. non-matching of prosody may be interactionally relevant to the participants and that it may have the power to drive sequential organisation of talk between two individuals.

The new direction of prosodic analysis in conjunction with interactional and gestural analysis suggested in this thesis tries to approach speech in a holistic form – or at least a bit closer to the holistic form than attempted previously.

Above, we have discussed previous research on the gestural aspect of conversation. In order to come closer to a holistic approach of analyzing face-to-face interactions it is necessary to take the visual modality including gestural analysis into account (Streeck, 2009), (Meyer, 2010).

In sum, the review of the literature on prosody and gesture from the different viewpoints (phonetics, gesture research and interactional phonetics) culminates in the following research questions.

2.6 Detailed research questions

RQ1: What are the sequential correlates of the social actions of alignments and non-alignments?

RQ2a: Are alignments performed with prosodic matches and are non-alignments performed with prosodic non-matches?

RQ2b: How can prosodic similarity be measured objectively?

RQ2c: What are the prosodic parameters that are responsible for the identification of prosodic matches and non-matches?

RQ3a: Are alignments and non-alignments performed with specific gestures?

RQ3b: Are alignments performed with gestural matches and non-alignments with gestural non-matches?

RQ4: How can the two modalities be combined in a sensible prosodic-gestural model?

The design of the study that should help to answer these questions is presented in the next chapter, introducing the material and the overall method.

3 Material and method

The cornerstone of CA is access to naturally occurring spontaneous talk, in which social actions can be identified. While studies within CA sometimes focus on a relatively small number of instances of a particular action or phenomenon, the collection of a substantial number of instances is useful in order to make the findings from such detailed qualitative analysis accessible for the wider research community. Additional to the requirements for CA, there are also requirements for the quality of the material of the corpus itself when a study makes use of automatic procedures for processing audio signals. While it is crucial for CA to analyse naturally occurring talk, from the perspective of a phonetic analysis, this involves all kinds of problematic effects such as overlapping talk, resulting in masking of the overlapping speakers. This causes a source separation problem, because acoustic analyses need clean signals which need to be attributable to the individual speakers. If only one signal is available, it becomes very difficult to separate the sources which originated from different speakers and differentiate them in separate signals. Therefore, despite recent advances in computational auditory scene analysis (CASA), (see Wang and Brown, 2006) a single audio stream is hardly viable, since it is extremely difficult to decide which bits of the signal belong to which speaker and which is only background noise. What seems to be a minor problem from the perspective of human perception becomes a major problem from a computational point of view. The separation of sounds subsequent to the recording stage will often remain as a persistent source of errors. Therefore it is a decisive advantage to have separate signals from separate speakers, if automatic processing of the signals is envisaged. Much of the data used in CA research is telephone speech, where the signals of the two speakers are separate by nature of the communication channels. However, most of these recordings involve only two participants who are visually isolated. Obviously, this makes a gestural analysis impossible, which is one of the prerequisites of this study. Things may change when video channels become more popular to be added to the audio channels for communication needs (e.g. “skype”), using the internet for combined voice and video streams.

3.1 Material

The research questions posed at the end of the previous chapter address two modalities (verbal talk with prosody and non-verbal gestures) from different angles (conversation analysis, interactional phonetics and automatic processing). This approach imposes many requirements on the material used. The corpus that was selected is described in Section 3.1.1. The orthographic transcription of talk is described in Section 3.1.2 and the gesture annotation is described in Section 3.1.3. The interactional environment of adjacent turn pairs from two speakers that is focused on is described in Section 3.1.4.

3.1.1 The AMI-meeting corpus

The AMI-meeting corpus (http://corpus.amiproject.org/) complies well with the requirements for the proposed study. It consists of round-table meetings recorded on multiple microphone channels and individual video cameras for each participant. It includes both, staged meetings (referred to as “scenario meetings”) and non-staged (spontaneous) meetings. The non-staged meetings are meetings that would have taken place anyway, as part of other research projects. Meeting participants include both native and non-native speakers of English. More information on the instrumentation in the meeting rooms can be found in the AMIDA final project report (2010) and in a summary by Carletta (2007).

As the AMI-meeting corpus provides audio channels from headset microphones, the effect of overlapping talk on the phonetic analysis could be kept to a minimum. Nevertheless, technical problems such as pop- and breath sounds on the close speaking microphones emerge, as will be

explained in more detail later. Some meetings have also been recorded on lapel microphones which are not exposed to pop sounds or breathing noise as much as the headset microphones, but these do not provide enough isolation of the talkers.

3.1.1.1 Meeting IDs

The individual meetings have identification numbers, for example "EN2009b". The first character stands for the recording location; here "E" for Edinburgh. The second character stands for the type of meeting; here "N" for naturally occurring. The series of four digits stands for a series of meetings at a certain location; here "2009" refers to the ninth series of meetings in Edinburgh. In the naturally occurring meetings, the final character is optional and is used for meetings "in the same series (i.e., of the same group, which may or may not be exactly the same participants). If there is only one meeting in the series, it could be omitted." (AMI, 2008); here "b" stands for the second meeting in a series.

3.1.1.2 Selection criteria for meetings

Several methodological factors limit the selection of data from the AMI corpus for use in the present study:

a) In order to reduce the impact of cross-speaker variability on the phonetic analysis, we selected meetings which involved a consistent set of speakers.

b) Meetings were chosen with participants who are all native English speakers, in order to reduce possible interference from the prosodic systems of other languages.

c) In accordance with the tenets of CA research, the selected meetings were naturally occurring and spontaneous rather than staged (scenario) meetings.

The meetings which were selected are designated EN2009b (51 minutes in duration), EN2009c (41 minutes) and EN2009d (85 minutes). In these meetings, researchers discuss software development and support for annotation of eye-tracking and language data, and how to use the data for subsequent analysis. Speaker A is a male computer programmer with a British English dialect; speaker B is a female data processing specialist with an American accent who lived in Edinburgh since 1988 (the recording was made in the year 2005); speaker C is a male postdoctoral psychologist with a Scottish dialect, and speaker D, who is only present in meeting EN2009d, is a female senior psychologist with an American dialect. Their corresponding identification numbers in the AMI-meeting corpus are MEE094 (A), FEE083 (B), MEE095 (C) and FEE096 (D). The selected meetings are quite specific in their organization and the constellation of the participants. In the first two meetings (EN2009b and EN2009c), the two male speakers A and C report on their software development progress to the female speaker B, who has a more senior position. One further senior scientist is present in the third meeting (EN2009d), in which a more open discussion evolves. Typically, one speaker produces stretches of talk, e.g. as a progress report, to which co-participants may respond. This reporting and discussing environment is comparable to the storytelling situation analysed by Stivers (2008) in some respects, for example in the asymmetry of contributions from participants.

3.1.2 Orthographic transcription

The AMI meeting corpus contains word-level orthographic transcripts, including start and end times for each word. While these transcripts provided an invaluable starting point, for our purposes it was necessary to re-transcribe relevant portions, some of which are presented below, using transcription conventions commonly used in CA research (cf. Appendix A).

Care was taken by the AMI transcribers to include all potentially relevant vocal tokens, such as laughter. In the transcription conventions it is stated that all speech and other vocalisations are transcribed verbatim, "as [they are] heard" by the transcriber (Moore, Kronenthal, & Ashby, 2005, p. 8). However, the AMI transcription conventions do not explain the relationship

between the orthographic transcript and the phonetic content of the utterance transcribed. For example, the phonetic basis for transcribing "uh" vs. "uh huh" (Schegloff, 1982; Jefferson, 1984) is not explained. We can only infer from the resulting transcripts what the phonetic properties of the different orthographic items (words) are. Similar difficulties have been reported by Ward (2000), who drew the attention to the importance of conversational grunts and presented phonetically inspired and "naive" transcriptions of those, resulting in tokens such as "um-hm-uh-hm". However, there were very few tokens of each type in his data.

A comparison of the orthography and the phonetic characteristics of individual tokens in the AMI data indicates that "uh huh" has mainly a mid-central vowel quality throughout (i.e. are more or less schwa-like), with an increase in air flow from the lungs through the glottis half way through the vocalisation. This may stop the pulsation of the airstream through the glottis or even cause frication [əhə], and can be heard as audible breathing. If the frication noise is less strong and the pulsation of the airstream does not cease, the vocalisation is perceived as having a non-modal voice quality in the middle: breathy [əə̤ə], creaky [əə̰ə], or voiceless [əə̥ə]. This splits the vocalisation into two parts that can be described as syllables. The end can also appear to be produced as voiceless [əhə̥], sometimes with frication noise coming from the glottis. Glottal stops are not generally observed at either the onset or the offset of the "uh huh". Different vowel qualities can also be found: e.g. more open, vowel quality ("ah hah" [ɐʰɐ]) or more fronted ("eh heh" ranging between [ə], [e] and [ɛ]). The token "uh huh" can be distinguished from other vocalisations that are based on a schwa-like quality, e.g. hesitation "uh", by its bisyllabic structure. Other tokens with two syllables may be nasal throughout: bilabial: "mm hm" [mʰm] or alveolar: "nn hn" [nʰn]. In the phonetic description here, we have turned special attention to the tokens "uh huh", "mm hm" and "nn hn", however it is assumed that the transcription of other "words" ("right", "yeah", "okay", "really", etc.) is similarly simplified. There are big steps from the phonetic diversity in vowels, and consonants to orthographic symbols. In this thesis, the segmental phonetics are not analysed for interactional differences between different turns – the focus is on pitch and intensity at the moment. Although theoretically it might be desirable to capture every phonetic detail, but this is not practically feasible for a large collection of data.

With the AMI transcriptions at hand, it was intended to move those transcriptions closer to a CA transcription. Therefore, the lexical representations of the "words" were used instead of transcribing in most possible detail the exact phonetic realisations – which would be incomplete anyway. We do this because the orthography covers all possible variation a reader or listener may ever have encountered. The same accounts for the transcription of the visual properties. All variation of the type "nod" is in the orthographic transcription "nod". Despite Kelly and Local's view on phonetic transcription and all its limitations due to the IPAA, it has to be put into perspective. This thesis does not study segmental features of the talk, for example segmental matching of aspiration, articulatory placement, etc. (which could be really interesting to do). With the current data, this is not possible. A huge amount of work would be required to get the corpus to a standard that made such analyses possible. Here, the information is made available at word level and the level of gesture types.

Simplifying the transcription does not mean simplifying the reality. By reducing the complexity from [ɪt] over /ɪt/ to "it", the last one contains all possible pronunciation variability that can be imagined or that a reader has already heard: for instance, with aspiration at the end [ɪtʰ], or glottal stop at the beginning [ʔɪt], or both [ʔɪtʰ], or [ɪʔ] and with a range of possible voice qualities and F0 movements. The rest is left to automatic processes on the acoustic record. Both are emphasised the qualitative interactional analysis on the one side and the quantitative acoustic analysis on the other side. The acoustic analysis is performed with automatic algorithms, lending the study objectivity. This way we arrive at two different, but complementary transcriptions, one is orthographic and one is acoustic. While Kelly and Local (1989) and most researchers in interactional phonetics do the acoustic transcription by hand, our

automatic processes have the potential of analysing much larger quantities of data and increase the reliability of results.

However there is some degree of flexibility in the CA transcription standards to incorporate phenomena like lengthening ("`:`"), aspiration (inbreath "`.h`" and outbreath "`h`"), etc. (see more on transcription symbols and conventions in the Appendix A).

Researchers of interactional phonetics argue that no phonetic detail can be ruled out a priori (Kelly & Local, 1989), in theory, but for practical analytical purposes, some may be ignored in the transcription and in the analysis. The current study is such an exception. Due to the requirements of the acoustic part, the transcription of prosodic patterns has to be omitted for the following reasons. First, the aim is to make acoustic analyses of prosodic features on data which is derived from interactional analysis. Because the interactional analysis is fundamentally based on orthographic transcripts, the availability of prosodic features in these transcripts would introduce circularity into the analysis steps. The conversation analyst should therefore be blind to prosody in this case. Second, even if no subsequent acoustic analysis was envisaged, the introduction of prosodic features in the orthographic transcription has the potential to bias the researcher's analysis, as many functions of specific prosodic contours have been proposed in the literature, which are still under controversial debate. A researcher would therefore be tempted to assume one or another function of specific prosodic contours without proper investigation of the conversational sequences at hand. The interactional analysis should be free of biased decisions. Therefore the material which could lead to such bias should be omitted. And third, the transcription of acoustic features, such as the prosodic ones is prone to errors. The agreement on the annotation of prosodic features, or prosodic categories, would be a research field on its own (Wightman, 2002).

3.1.3 Gesture annotation

The AMI meetings are face-to-face interactions. One aim is to investigate for the present data, if the presence or absence of nods (and other gestures) is related to alignments and non-alignments. To avoid making assumptions, prior to the analysis about the role of particular gestures, all major gesture types that are observable in the selected meetings have been annotated for the current study. Only the scenario meetings were annotated by the AMI team with so-called "actions" or "events". The non-scenario meetings, which are used here, were not previously labelled for gesture.

The annotation instructions given to the AMI labellers have merged action or head event recognition and head movements. They make rough distinctions between deictic (pointing) and non-deictic hand events and make rough distinctions for trunk gestures. Unfortunately, "*reliability test results are not currently available for this scheme*", as it is stated in the AMI transcription conventions (Moore, Kronenthal, & Ashby, 2005). But some changes need to be put in place for the gesture annotation to make it more useful for the gesture analysis intended in this study. On the one hand it has to be more detailed and include highly communicative facial gestures such as raised eyebrows and frown. On the other hand it needs to make some simplifications, as "fiddling with an artefact" or "fiddling with the hand" are both hand gesticulations and both may be communicatively relevant. Because not all kinds of hand gesticulation can be labelled distinctly and consistently, the generic term *hand gesticulation* is used in the current study to accommodate all kinds of variation. For the trunk gesture layer, the *shrug* is distinguished from other body movements. Because postures such as sit upright, lean forward, lean backward are necessarily presented throughout the conversation, a turn-by-turn analysis, as it is envisaged in this study would be predominated by this type of gesture, therefore it is not annotated here in this way. However, immediate trunk movements are recorded as *body movement*. Such movements are for example the movement from leaning forward to leaning backward, or repeated sideward movements or movements with the shoulders which cannot be assigned to a *shrug*.

In summary, the major gesture types that have been observed in the selected non-scenario meetings are: *hand gesticulation*, *body movement*, *nod*, *blink*, *head shake*, *shrug*, *raised*

eyebrows and *frown*. Because blinks and nods very often co-occurred, an additional category *nod and blink* was used to account for this.

In our annotation of the selected meetings, only the major gestures have been recorded. That means if there was a large hand gesticulation and a hardly perceivable frown co-occurring with the hand gesticulation, only the hand gesticulation was annotated. This rule was followed in order to avoid multiple layers of gesture transcription where it is hard to draw a line between motions that are gesture and other motions. Because of the high complexity of gesture motions (direction, amplitude, etc.) it was only annotated if a gesture has been present or not (cf. Stivers, 2008). In future research it may prove valuable to record the motion of gestures in more detail (cf. Whitehead, 2011). The annotation of gestures was done through inspection of the videos. Each meeting was video-taped with six cameras: one in a corner of the room, one placed overhead and four individual cameras, one for each participant who is captured from the chest upwards.

The annotation software ELAN was used (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006). By visual inspection it was judged whether a speaker's head movement was a nod or not, whether it is a passive movement that can be attributed to body movement, whether the motion of the head and shoulders was a shrug or not, whether the body movement was due to the movement of the trunk or the hands, whether the motion of the hand was gesticulation or just grabbing an object. Whenever a gesture ceased and a new gesture started – and may it be from the same type – each gesture was assigned a separate annotation segment.

3.1.3.1 Facial displays

Most of the authors discussed in Chapter 2 Section 2.2.5 look at individual gestures and at the meaning they may have – alone or in connection with speech. With regard to facial displays, Bavelas and Gerwing (2007) state that "it is possible to analyse faces reliably, objectively, and quantitatively in terms of their conversational meaning (Chovil, 2005) and that experimental manipulation can reveal the functions of facial displays in dialogue. For those who are intrigued by facial actions that are timed to conversation rather than emotion, this is an unexplored and promising frontier." (p. 302).

Blink

We do blink all the time to keep the eyes from drying. This movement of the eye lids occurs every now and again and it is a very quick movement. Different intervals between eyeblinks are reported in the literature and range between 2 and 10s or 2.8 and 4s. Different spontaneous eyeblink rates (SEBR) were reported which range from 12 eyeblinks/min to 15 eyeblinks/min, as summarised by Doughty (2001). The main concern is that these numbers have been reported without reference to "the conditions under which such inter-blink intervals or eyeblink rates were obtained" (p. 712). Considerable differences have been found between reading and conversational conditions, with median values of 7.3 eyeblinks/min and 23 eyeblinks/min respectively. Example observations of eye blinks in Doughty (2001) show relatively constant intervals between eye blinks in the reading condition, but in the conversational condition, blinks sometimes come in clusters. Observations in the meetings of the AMI corpus also indicate that these clusters of blinks may co-occur with speech. Similar to the reasoning on postural position (forward, backward, upright), we are not interested in all kind of spontaneous eyeblinks which occur necessarily, but only in those which have more than the purpose of watering the eyes. There are also long lasting eyeblinks, which are not the abrupt movements of closing and re-opening as in endogenous eyeblinks. A third version of eye-lid movements can be observed, where the eyes close half way and the eye-lid jiggles without complete closure. Blink clusters, long lasting and jiggling eye-blinks are recorded and assigned a *blink* tag.

Eyebrow movement

The movement of the eyebrows is another sort of facial display that we consider in our annotation of gestures. They can mainly move in two directions: upwards and downwards, however, they are coupled with other muscles of the front which can pull them together. The resulting gesture can be described as either *raised eyebrows* or *frown*. In Chovil (1991), these two eye-brow movements are called "brow raising [and] lowering" (p. 169).

Raised eyebrows is an upwards movement of the eye brows, which sometimes results in the skin of the front folding up horizontally. The eyes are usually open. If they are closed or half closed, the eyelids become exposed.

A *frown* is the opposite movement to the eyebrow rise. The eyebrows are lowered and sometimes pulled together towards the middle. This often results in vertical folding of the skin of the front between the eyebrows. Then the eyes are less visible, the eye-lids become hardly visible.

3.1.3.2 Hand gesticulation

All movements of the hands which are not oriented towards a specific task, e.g. grabbing a pen or scratching the skin, are annotated as *hand gesticulation*. They are comparable to what Kendon (1982) specifies as "gesticulation", however it is not necessarily the case that the hand gesticulation has to coincide with a verbal utterance. Although hand gestures have widely been separated into certain phases, such as preparation, stroke and retraction phase, these phases have not been distinguished here. The beginning of the hand gesticulation tag was set to coincide with the start of the preparation phase and the end of the tag was set to coincide with the beginning of the retraction. If the hand does not retract, but restarts gesticulation (similar or different to the previous one), a new segment starts at that point.

3.1.3.3 Shrug

A *shrug* is a bodily movement where one shoulder or both are raised. Often, the shoulder movement coincides with a sidewards movement of the head towards one of the shoulders (if only one shoulder is risen it is typically the one shoulder towards which the head moves). No phases of a shrug are distinguished.

3.1.3.4 Head movement

The shrug – although it sometimes involves motion of the head – is distinct from other head movements. Our focus is here on head *nod* and *head shake*.

Nod

A vertical flexion of the neck is annotated as a head *nod*. It can be a single upwards or downwards movement or a repeated movement, like a pendulum. We do not distinguish between degrees in strength or count how often the head went back and forth until the movement ceases. Once the nodding gesture restarts, a new segment is used to contain the new head nod. This is in line with the annotation procedure used by Stivers (2008), who considers

> *"all evidence suggests that nods (whether one or multiple, whether deep in their vertical trajectory or shallow, whether rapid or slow) are still tokens of a single gesture type" (p. 37).*

Head shake

Horizontal turns of the head are annotated as *head shake*. Here, the head rotates left and right around the upper cervical spine. Similar to the head nod, this movement can be repetitive and of

different strength and amplitude, though this is not recorded for the purpose of the present study.

3.1.3.5 Body movement

We chose to have one more gesture category for movements of the body, which are neither head nor shoulder nor hand related. All of these body parts are attached to the torso and are relatively free to move. However, if the torso moves, the head almost inevitably moves with it. Conversely, if the head moves, the rest of the body also seems to move. It is therefore important to identify the origin of the movement. If the movement of the head and shoulders cannot clearly be assigned to either head nod, head shake or shrug, it falls into this category *body movement*.

3.1.3.6 Summary of gesture annotation

Unlike the annotation system which was used by Chovil (1991) to develop linguistic categories, our annotation system should help to develop interactional categories. Our system is designed to aid the analysis of specific gesture types in relation to specific social actions and the comparison of adjacent gestures in relation to specific social actions.

For our purposes, it is necessary to compare whether the gesture type used by one speaker at a certain sequential position is the same as the gesture type at a certain different position in that sequence. This way should offer a possibility to uncouple the gesture from its assumed function and relate it to the gestural and conversational environment.

With this approach, we also introduce a slight simplification. Because we do not record differences between, say, a small and a large head shake, we will still match head shake with head shake. However, a similar simplification has been made by other authors, e.g. Stivers (2008), who did not distinguish between small and large head nods.

3.1.4 Turn organisation and IPs

In order to answer the question whether talk from one speaker aligns or disaligns with the prior talk of the prior speaker, the utterances of the two speakers have to stand in close relation to each other: there must be at least a first utterance (domain; from the prior speaker) and a second utterance (target; from the target speaker). The reaction which happens after these two utterances, i.e. the treatment of the target, tells us how the target utterance has been interpreted by the prior speaker of the domain utterance (and the other speaker(s) that may come after a target).

According to the research questions (RQ1 and RQ2a), there are two main hypotheses: The first is that the one speaker aligns or disaligns with the ongoing course of action of the other speaker. The second is that this aligning and disaligning action depends on the degree of the prosodic match of the two turns. In order to make such a comparison, it is necessary to define the units or components of utterance that are to be compared. The prosodic-acoustic analysis (described in Chapter 5) shades into the selection process of the turns which are to be analysed interactionally. The acoustics are not analysed continuously over the whole duration of the signals, but need to be provided with information on the regions of the signal that should be analysed. It means that the components and their boundaries are well defined. There are several possibilities for such components (see Section 2.3). Such components can be the turn, the turn constructional unit, the “chunk” or “turn constructional phrase” (Szczepek Reed, 2010b) or the “sequential slot” (Szczepek Reed, 2012a, p. 26). This is discussed shortly in the following.

A turn can consist of several turn-constructional units, and a turn-constructional unit can have several prosodic accents or tone groups. This means that the answer to the question, which part of one utterance should be compared with which part of another utterance, becomes difficult. A unit is needed that embodies the notion of a component that is distinct from its surroundings, i.e. neighbouring units. It was decided to take the intonational phrase (IP) as the basic component,

as it carries only one phrasal accent. Therefore the intonational phrase is less confusable across and within the different speakers' utterances.

This means that the adjacent turns of the two speakers have to be analysed according to their intonational phrase structure first. At this point, the reader might get the impression that this prosodic analysis may shade into the interactional categorisation and may cause a problem of circularity, as the aim is to compare the utterances for their prosodic characteristics at a later stage. The intonational phrase analysis is expressly performed to ensure that comparable units can be compared later. No other prosodic analysis is performed at this stage.

Looking at the transition from the first speaker to the second, there are several possibilities for sequential organisations: The first speaker can be followed by a single IP from the second speaker, before the first speaker regains the turn. An example for such a turn organisation is the "uh huh" of speaker B in Extract 1 or the "by" in Extract 2. They are both produced with one single intonational phrase. But in other cases, the second turn may contain several IPs. If this is the case, then each IP is addressed separately. For example where a turn consists of two intonation phrases, one can speak of a first component and a second component of the turn. The second component of a turn is then not directly adjacent to the last IP of the prior speaker's turn. An example of such turn organisation is the turn "uh yes with the um" of speaker C in Extract 13 (see Chapter 4), where the first component "uh yes" is produced with one intonational phrase, while the rest "with the um" is produced with another intonational phrase. There is no gap following "yes" to suggest from the transcript that there are two components here. Therefore a prosodic analysis into intonation phrases was necessary when compiling the collection of instances of adjacent turn pairs.

In the following, the term *target* is used for the second speaker's IP, because it is the IP under question. It means that the aim is to assign it the interactional label. The first speaker's turn – or its final IP – is the *prior turn* or the *domain*. The turn that follows the target is called *successive* utterance or IP. Everything after that is called the *subsequent* utterance, or turn. Note that a gesture can also form a turn.

It will be seen in Chapter 4 that the second turn can be a single word / token ("yeah", "okay", "right", etc.) or a sequence of words ("oh yeah", "they are the paid", "it's not too bad", etc.), but its scope is limited to the duration of one intonational phrase for the purposes of the prosodic analysis in Chapter 5. The target turn can – but does not have to – overlap with talk from the prior speaker. Such overlap can be partial or complete. The second speaker can overlap with the continuation of the prior speaker or the prior speaker can overlap with the second speaker. The target turn can be followed by a pause, inbreath, outbreath, hesitation, or none of those. Target turns can be surrounded by gestures (from the same speaker or the other speaker(s)). Gestures can appear anywhere in the sequence. Therefore gestures can overlap with other gestures and they can overlap with verbal expressions. Gestures can substitute verbal expressions. They count equally as a turn.

In brief, the following organisation patterns were found to make such an interactional categorisation into alignments and non-alignments possible:

- Target turn in the clear (no overlap)
- Target turn in partial overlap
- Target turn in full overlap (e.g. choral co-production (cf. Lerner, 2002)).
- Target followed by more talk from the prior speaker
- Target followed by short silence and then by prior speaker's talk
- Target followed by short silence and then by talk from the same speaker
- Target followed by short silence and then by talk from a third speaker
- Target speaker continues without break

There are several reasons for analysing such short utterances. First, there is evidence in the literature (Gardner, 2001), that short utterances of the same lexical shape, such as "uh huh" can perform different actions, depending on their contextual (and possibly "con-prosodic") use. It means that the interactional meaning of an utterance is not predetermined by the lexical or syntactic meaning alone.

Second, the analysis of adjacent turns limits the duration of the turn components to one intonational phrase. This makes it possible to compare across speakers according to their prosodic similarity without the risk of confusing similarity of non-related IPs.

The method of how we can derive the interactional meaning from these turn components or IPs at second position will be demonstrated in Chapter 4 using close analysis of these components and their immediate environment. This includes that we should investigate the content of each part of the sequence, i.e. investigate how the participants build the sequence of first, second and third turn. We demonstrate the interactional consequences which the sequential structure has on the conversational actions.

3.1.5 Acoustic preparation

In order to address the prosodic matching hypothesis, reliable methods are needed to measure the acoustic similarity of the target turn and the immediately preceding turn. Since F0 and intensity are likely to be prominent factors in the listener's perception of a prosodic match (Szczepek Reed, 2006), we start by comparing the F0 contours and the intensity of the adjacent turns. Although there are many more features which can be considered to be relevant, it was decided to take those parameters, as they can be used in an automatic comparison (Kousidis, Dorran, McDonnell, & Coyle, 2009). Other features, including speech rate, suggested by Szczepek Reed (2006) might also be possible to compare, but this would require further research.

The first step in the acoustic study is to normalise the F0 range and the intensity range, in order to deal with cross-speaker discrepancies and channel differences, as shown in other studies on prosodic matching (see Section 2.2.3). The second step is to compare the F0 contours in order to quantify their similarity.

As a further refinement, with the aim of more closely approximating participants' perception, the different F0 contours are compared and then the resulting similarity is weighted by their intensity. This last step is not crucial for the general methodological approach used in this thesis, but it constitutes an approach towards a more holistic analysis of prosody. As the literature suggests, F0 and intensity are not strictly divisible and are not independent from each other (Neuhoff & McBeath, 1996).

3.2 Method

First, a set of adjacent turns that meet the requirements of organisational patterns are collected (turns in the clear, in overlap, followed by the prior speaker, followed by the same speaker, etc.).

Second, the turns are classified into the interactional categories of alignment and non-alignment according to the sequential organisation, namely the orientation of the interacting participants towards the target turns. The prosodic resources are ignored for that task. The gestural resources of the target turn are also ignored; only the gestures of the prior and the post turn of the prior speaker are analysed. How far the sequential detail discovered in the interactional analysis can be recognised by a second analyst is evaluated by measuring inter-rater agreement.

Third, an acoustic measurement for prosodic similarity is devised in order to investigate the prosodic matching hypothesis. It is tested statistically if the aligning turns are more similar to the prior speaker's turn than non-aligning turns.

Fourth, the alignments and non-alignments are analysed according to the gestural cues which accompany them. Two hypotheses are tested: a) alignments and non-alignments are consistently produced with specific gestures (alignments with different gestures than non-alignments), and b) alignments and non-alignments are distinguishable through gestural matching and non-matching.

Fifth, the acoustic and gestural resources are compared. It is tested how the prosodic properties (matching and non-matching) are related to the gestural properties.

4 Interactional analysis

One substantial undertaking in this thesis is to establish a collection of instances of adjacent turns, which can be distinguished along the lines of the social actions of alignment and non-alignment. The research question that should be answered is what the sequential correlates of alignments and non-alignments are (RQ1). It is important to show that this distinction can be grounded in the orientation of the participants to each others' talk. In this chapter, the interactional analysis of adjacent turn sequences is described. It culminates in a categorised collection, ready for prosodic and gestural analysis.

4.1 Sequential organisation of alignments and non-alignments

This section demonstrates the procedure used to collect adjacent turn pairs which have the potential of exposing aligning or non-aligning actions. The following examples show that the orientation of participants can help to interpret the interactional context of adjacent turns in order to make predictions about the social action performed in the sequence. The observations demonstrate that patterns exist in the organisation of talk and gesture.

In Chapter 2, Section 2.1, some of the resources were discussed, which conversational participants might employ in order to perform aligning and non-aligning actions. The resources to be examined here are the organisation of overlap, inbreaths, outbreaths, hesitation, pauses and gesture organisation of the *prior* speaker. Prosodic resources (F0, intensity, tempo, voice quality, etc.) are dealt with in Chapter 5 and the gestures of the *target* speaker in Chapter 6. However, in this chapter, the gestures of the prior speaker are analysed. As a short reminder, the "prior turn" or "prior speaker" is the first turn/speaker in an adjacent turn pair. The "target turn" or "target speaker" refers to the second turn/speaker in that adjacent turn pair. For this turn, it should be evaluated whether it is aligning or not with the prior turn/speaker.

Therefore it is necessary to analyse what the agenda of the prior turn is, i.e. what the participant is doing in that turn. In accordance with Szczepek Reed's notion of "action trajectory" (2012a, pp. 13, 18), "previous course of action" (p.14), "previously established interactional project" (p.16) and "previous conversational activity" (p. 20), the term *prior agenda* is used to refer to the action that a participant is pursuing. It has been demonstrated that these prior activities can be described consistently. For example, in radio phone-in-programmes, the moderator and hearer routinely introduce each other and start with a greeting sequence as in an extract from Szczepek Reed (2012a, p. 20):

1. BA: TIM.
2. good MORning.
3. TI: <<h> ↑HI ↓BARbra,>
4. (0.12)
5. BA: <<h> ↑HEL↓lo TIM,>
6. TI: haven't tAlked to you in a LO:::NG TI:ME.

Here, Tim is a caller who is selected by the moderator Barbara and first called by name (line 1) and greeted with "`good morning`". The activity done by Participant BA is doing a greeting. It is demonstrated by Participant TI that the activity done in the prior turn is a greeting by greeting back ("`hi barbara`"). Although Szczepek Reed does not explicitly interpret the "course of action" as greeting, this extract shows that the activity trajectory that is outlined in

line 1 and 2, with the initial greeting and summons ("tim good morning") is identifiable as part of a greeting sequence (see detailed interpretation of the entire extract by Szczepek Reed, 2012a, p. 25).

Another identifiable action that is performed in a "prior turn" is taken from Szczepek Reed (2012a, p. 22):

1. BA: well you DO have two chIldren an-
2. [.hh
3. MI: [<<l> YEAH.>
4. BA: <<breathy+p+h> it SOUNDS like um;
5. (0.18)
6. you're a ↑GOOD ↑DAD.>
7. (0.31)
8. MI: <<breathy+p+h> ↑Oh ↑THANK you.>

In line 4 to 6, Participant BA is making a compliment ("it sounds like um your're good dad"). At least that is how it is picked up by Participant MI: "oh thank you". Thus, the action trajectory of the prior turn can be identified as making a compliment. That is what is meant by the term that will be used in the following: the prior speaker's "agenda", "action" or "project". Besides greeting and making a compliment, other action trajectory formats of prior turns that were identified by Szczepek Reed are: asking a question, making a proposal, repair initiation, first claims or news delivery, display of appreciation.

The observations that follow show that the prior speaker has a tendency after a turn of a second speaker to either continue on the prior agenda or to orient towards the second speaker's turn in a different way, as if it is moving the talk in a new direction. If none of these possibilities are chosen by the first speaker, a pause may develop, after which any speaker can take the floor, or the second speaker starts immediately a new agenda (with a new intonational phrase). This organisation is not limited to the verbal part of the orthographic transcript but can also be observed in the gestural part. For the production of a gesture in the prior speaker's turn, the term "gestural agenda" is introduced here in analogy to the prior speaker's verbal agenda. After the target turn, the prior speaker can either continue on the prior gestural agenda or start a new one with a different gesture to the prior one, orienting to the target turn as a turn that needs specific treatment.

The focus of this section is the positioning of the adjacent turns in relation to each other and to their immediate verbal and non-verbal context. The target turn can for example occur in the clear or in overlap, it can be surrounded by stretches of silence, it can do collaborative completion or sometimes it can be made up of two components, which perform distinct actions. How these environments work together to influence participant alignment is discussed in order to answer the research question (RQ1).

4.1.1 New developments in transcription

The transcripts of Extract 1 and Extract 2 are displayed in the traditional CA transcript style. This gives every turn of a speaker a separate line and marks short turns which overlap with another speaker's turn with square brackets. For the two extracts above (see Section 2.1.3), this way of displaying orthographic content, and even the visual aspects of the interaction, enables the reader or analyst of the transcript to follow the interaction chronologically, i.e. who is speaking, gazing and gesturing when.

However, there are instances when the chronological order is not clear anymore. This happens when some talk or gesture overlaps over more than one line or when one overlapping region is followed by another overlapping region. It becomes unclear to which part the overlap belongs: to the upper (earlier part) or the lower (later part).

There might be some techniques which could resolve this problem, for example brackets or braces which extend over several lines and which bundle overlapping regions and separate them from other overlapping regions, or other separating lines. But no technique has been developed yet for displays on computer screens or print. In a discussion of visual representations of interactional transcripts, Schmidt (2003) is concerned with how to map elements from a logical structure (as in an XML document) onto elements of a visual structure (as in an HTML document). Schmidt suggests displaying them in orthographic transcripts in the format of a musical score where the chronological order is clear without ambiguity. An example of such a visual representation of temporal structure and other levels of linguistic analysis can be found in Schmidt (2010), p. 27. All transcripts from this point onwards are presented in that score format.

In particular, gaps are marked in the way that pauses are marked in traditional transcripts. Whenever no verbal utterance is produced by any speaker, the duration of that time interval is displayed in brackets in the bottom part of the current line, e.g. (0.5) is gap of half a second where no participant produces any verbal utterance (see line 2 of Extract 3).

4.1.2 Target turn in the clear

Many of the target turns are produced in the clear, i.e. without verbally overlapping with, or being overlapped by talk from the prior speaker. The "`yeah`" in Extract 3 is an example of a turn organisation where the target IP, here a complete turn from Participant B appears between two extended turns from Participant A.

Extract 3: "drift correction" (B-03:27)

[1]

Gest_A		*body movement*		*body movement*	
A	but it loo* it looks like it was doing all the	calibration	and	drift correction things	that s*

[2]

Gest_A		*head shake*	*shrug*	*body movement*			
A	ellen	was wanting on	it so	°hh		so that	don't think there's anything ...
Gest_B					*nod and blink*		*nod*
B					yeah		
Gest_C				*nod*			*nod*
				(0.54)			

[3]

A	to add to the software (-) for that part ...

In line 1 and the first half of line 2, Participant A has an extended turn, reporting on work about software development, which can be considered as his project. Part of the verbal utterance is accompanied by body movement gestures. Towards the end of that turn there are a head shake, a shrug on "`so`" and more body movement during an inbreath. Participant B's "`yeah`" is accompanied by a nod and a blink. It is immediately followed by further talk from Participant A starting with "`so`" and continuing his prior reporting on the software development work. This continuation is thus still on A's original agenda. In the talk that follows the target, Participant B's "`yeah`", Participant A continues without any other interruption. The fact that A's continuation on his prior agenda is not treated as problematic by B is more evidence for the interpretation that B's target turn was not designed to initiate a new action or to introduce a new agenda. It suggests that the target turn is an alignment.

4.1.3 Target turn followed by silence

In Extract 4, which is a direct continuation of the first two lines of Extract 3, the target turn "`mm-hmm`" from Participant B is not followed by a long turn from A, the prior speaker. Only a hesitation "`um`" follows the target and a stretch of silence of 0.55 seconds develops. Participant

A does not produce any talk that could give evidence that his prior agenda (reporting on software) is continued. Here, Participant A chooses not to continue on prior talk even though Participant B allows room (the "um" from Participant A plus 0.55 seconds) for a possible continuation. After this opportunity has passed, Participant B takes the floor, cuts off the prior agenda (on "software") and starts with another topic ("so i've got a new (.) um trainee ...") progressing with an agenda that is related to talk that has come before the topic of Extract 3 and Extract 4, which was that Participant A wanted to run a test recording with a random subject. From the extracts alone it is not entirely clear that the turn from Participant B here refers to that previous topic. However, the interaction continues with B's suggestion to borrow the trainee for this test recording.

Extract 4: "software" (B-03:34)

[3]

Gest_A		*nod*				
A	to add to the software (-) for that part			um		
Gest_B			*nod and blink*		*blink*	
B			mm-hmm		so i've got a new (.)	
Gest_C			*nod*			
		(0.38)		(0.2)	(0.35)	

[4]

Gest_B		*body movement*	
B	um trainee who might need amu	sement at	some point (.) ...

Looking at the sequential organisation in more detail, it stands out that after Participant A comes to the end of his turn, a short stretch of silence (0.38) develops. With the end of the prior turn and the start of this short silence, A starts nodding and continues nodding through B's "mm-hmm", through his own "um" and through another silence of 0.2s. It can be argued that both, the verbal and the gestural agenda, are continued after the target turn "mm-hmm", the gestural agenda by continued nodding and the verbal agenda with the hesitation "um".

After the target turn, the hesitation from Participant A and the stretch of silence, which constitutes a point where Participant A has the opportunity to continue talking on the prior agenda, this opportunity elapses with the start of Participant B's turn "so i've got a new ...". Participant B is not interrupted and does not interrupt her talk subsequently, except for the micro pause "(.)".This indicates that the prior speaker leaves the floor that was initially re-allocated to him by the "mm-hmm" to Participant B, who takes up that opportunity to re-introduce a previous topic.

This transition from one speaker to another is negotiated over the aligning turn "mm hmm" that is not interrupting the prior speaker's agenda. It leaves the opportunity to the prior speaker to continue on his agenda. This is done on both the gestural and the verbal side. More examples, where the gesture of the prior speaker during and after the target turn is analysed, follow below (See also Section 4.2.2).

An example where no verbal, but only gestural continuation of the prior speaker follows the target turn is contained in Extract 5. Here, the target can be both the "yep" of Participant A and the "yeah" of B, as they occur almost simultaneously. These turns are followed by silence (0.87). But Participant C starts nodding during in- and outbreath well before (0.85) Participant A starts the "yep" and Participant C continues nodding for 0.42s even after the end of the "yeah" from B. This indicates continuation on the prior speaker's gestural agenda, suggesting alignment of the target turn(s). It is only after the opportunity to add more talk to the prior agenda by Participant C has passed that Participant B takes the floor, starting with blinking and "although if ...", which is disagreeing with C's talk, and thus initiating a new agenda afterwards. Therefore, the "although if" could be classified as non-alignment.

Extract 5: "reasonable" (B-07:53)

[1]

Gest_A	*nod*
A	yeah yep
Gest_B	*nod*
B	yeah
Gest_C	*nod*
C	a reasonable target uh for the for the tracking purposes °hh hh°
	(0.85) (0.16) (0.26) (0.07)

[2]

Gest_B	*blink*
B	although if if you had a big thing and it was jiggling around like anybody's business
	(0.38)

[3]

B	like real eye track traces ...
Gest_C	*nod*

4.1.4 Target turn in overlap followed by silence

A similar case, however in overlap, is found in Extract 6, where the "yeah" of Participant B (end of line 1) is followed by silence (0.54) and subsequent talk "i mean" from the same speaker. The stretch of silence is important, as it leaves space for the prior speaker to continue speaking. The "i mean", accompanied by raised eyebrows (beginning of line 2) does not start before the prior speaker has let pass this opportunity.

Extract 6: "moved over" (B-00:03)

[1]

Gest_B	*raised eyebrows*
B	((sound: laugh)) okay yeah i mean
Gest_C	*hand gesticulation body movement*
C	uh yeah so we've moved over and (.) on to the other eye link so uh
	(0.07) (0.54)

[2]

Gest_B	*raised eyebrows*
B	m* my view on this was (0.14) that um: the need for ...
Gest_C	*raised eyebrows*
C	'cause she's running at the moment

Considering the "i mean" as target (it is not really adjacent, but close to the prior talk of Participant C), it projects a next action of a certain type, for example an assessment, as in extract 5 in Goodwin and Goodwin (1987, p. 21). It constitutes the beginning of a new agenda, i.e. what Participant B's opinion or "view" on the affair was. Therefore she would have the floor (right to speak) afterwards. However, Participant C produces an increment "'cause she's running ...", orienting to the "I mean" as a turn that shifts the agenda away from his agenda (display of his point of view). He treats the "I mean" as a turn that requires more explanation. Here, Participant C chooses to specify the reason for why they have "moved over and (.) on to the other eye link". An increment would have been allowed by the "yeah", however this continuation comes too late, as Participant B has already initiated her own agenda. This causes Participant B to defend the floor and overlaps Participant C's increment with a truncated "m*", the beginning of "my view on this was ...". (See French and Local (1983) and Kurtić, Brown and Wells (in press) on turn competitive and

non-competitive overlap.) The two different sequences that the "yeah" and the "I mean" provoke, indicate that the "yeah" allowed the prior speaker (C) to continue on his prior agenda whereas the "I mean" initiated a new agenda. If the "I mean" was designed as allowing the prior speaker to continue on the prior agenda, the continuation of the prior speaker on his agenda would not have caused Participant B's floor-defending actions while that continuation of Participant C is in progress. Finally, this even causes Participant C to give up on giving the reasons for his view in the middle of the increment. Therefore, this example (the "I mean") represents a non-alignment. The "yeah" represents an alignment.

4.1.5 Full overlap

There are cases, where the second speaker overlaps the talk of the other speaker, even before a transition relevance place has been reached, as can be seen in Extract 7. After the short turn "okay" of Participant B (line 1), the prior participant, A, continues on the prior agenda which was to explain a technique for identifying the location of a specific sound. However, the "okay" is not ignored, as after the overlap A repeats "exactly", the overlapped part.

The "perfect" of Participant B (line 2) also overlaps A's talk ("now °h") before he (Participant A) comes to a transition relevance place. Again, this overlap is not treated subsequently as an initiation of a new agenda, as A continues on his prior agenda without orienting to the "perfect" in any "specific way". Additionally, Participant B does not add more talk after the "perfect". If she *did* continue, e.g. by defending the floor (as in the example "m* my view on this" in line 2 of Extract 6), this would indicate that she is treating participant A's talk as a challenge for the floor. (The notion "specific way" is repeatedly used in the following to refer to an orientation of the prior speaker to the target turn requiring special "treatment" or "work" from the prior speaker. This is in contrast to the orientation of the prior speaker to the target turn "allowing continuation on the prior agenda".)

It was discussed above (on Extract 4) that sometimes the organisation of gestures from the prior speaker can resemble a continuation or a non-continuation of a gestural agenda, as it often coincides with the continuation or non-continuation of the verbal agenda. In Extract 7, this can also be observed.

Extract 7: "really clear" (B-06:19)

[1]

Gest_A		*body movement*		*head shake*		*hand gesticulation*	
A	°h	did it with ten kilo-her	tz and it's just	(.) re	ally clear	(.) where you can tell	exactly
Gest_B							*nod and blink*
B							okay
Gest_C					*nod*		

[2]

Gest_A						*body movement*			*shrug*	
A	where	exactly	where	the beep is	now °h		so	y	eah so tha*	it's ju* just a matter of...
Gest_B				*nod and blink*		*nod*				
B					perfect					
Gest_C						*nod*			*nod*	
						(0.19)	(0.31)			

Participant A's hand gesticulation in line 1 extends before, during and after the "okay" from Participant B (into line 2). This gesture continuation coincides with the verbal continuation, suggesting alignment of the target turn. Regarding the turn "perfect" this is less clear. The body movement of Participant A does start with the end of Participant B's "perfect". However, Participant A has reached a place of syntactic completion, which makes turn transition possible. Being handed back the floor directly by Participant B's turn, Participant A searches for a possible continuation of the prior talk.

4.1.6 Partial overlap

An example with partially overlapping turns is in Extract 8. Participant C comes to a possible transition relevance place (end of line 1), as the syntax indicates. During the following inbreath (0.79s), the gesture changes from hand gesticulation to body movement. The "oh really" from B and the continuation "there was like" of Participant C overlap. With the end of the "oh really", Participant C interrupts the turn that he has just started and says "yeah mm-hmm". This break in the course of Participant C's talk indicates that he orients towards the "oh really" as a turn that needs special treatment.

The gesture of Participant C just after the "oh really" is different to the gesture before the "oh really". The prior gesture is a body movement and the following gesture is a nod. This non-continuation of the gesture indicates that the target turn is oriented to by this nod. The "oh really" was treated to need some extra "work", suggesting the target turn is a non-alignment.

The interaction in Extract 8 proceeds after Participant C's orientation towards the "oh really" from Participant B with the "oh okay" that partly overlaps with the "mm-hmm" of Participant C and is followed by a pause. The prior speaker has the possibilities to continue, to add more talk on the prior agenda or not, or to show orientation to the "oh okay" from Participant B. None of this happens and the third speaker (A) takes the floor.

The gesture of Participant C after the "oh okay" is the same as during the "oh okay". Both are body movement. It indicates that continuation on the gestural agenda was permitted, suggesting alignment of the target turn.

Extract 8: "eye" (B-07:09)

[1]

Gest_C	*body movement* *hand gesticulation*
C	ellen wanted something to look more like an eye (.) so that you knew that was the eye °h
	(0.44)

[2]

Gest_A	*hand*
A	yeah sos oso we've got like uh a
Gest_B	*nod and blink* *raised eyebrows* *nod and blink*
B	oh rea lly oh okay
Gest_C	*body movement* *nod* *body movement* *body movement*
C	hh there was like yeah mm m-hmm
	(0.35) (0.4)

[3]

Gest_A	*gesticulation*
A	big circle with a ...

Another example with partially overlapping turns is found in Extract 9, which shows subsequent orientation of the prior speaker towards another speaker's turn. Participant C comes to a transition relevance place (beginning of line 2), as the syntax is complete and the speaker breathes in. At this point both speakers overlap; Participant B says "just the parallelogram" and Participant C starts to continue, which could be on the prior agenda, as he uses a connecting pronoun ("it") which may refer to the "it" in the prior talk. After B's "just the parallelogram", Participant C first says "yeah" and then recycles B's utterance: "yeah just the parallelogram". After the following inbreath, hesitation "so" and silence (0.76), Participant C continues on his prior agenda (on rotation and vertices).

Extract 9: "mirror image" (B-09:18)

[1]

Gest_A	*nod*
Gest_C	*hand gesticulation* *hand gesticulation* *blink*
C	just totally flipped uh in in a mirror im[a]ge not rotate dso it wasn't anything to do with

[2]

Gest_B	*body movement*
B	just the parallelogram
Gest_C	*hand gesticulation* *nod*
C	rotation °h it was f* yeah just the parallelogram °hh so
	(0.33) (0.76)

[3]

Gest_C	*hand gesticulation*
C	uh entirely sure must be some way of how the vertices were described (.) or (.) ...

The stretch from the beginning of Participant B's turn until the end of Participant C's recycling of that turn represents an insertion sequence. B is checking her understanding that the topic is "just the parallelogram". At least Participant C treats it as an understanding check: like a turn that needs some work from him, e.g. some clarification. With the orienting "yeah" and the recycling of the other participant's turn, the prior agenda is interrupted, i.e. not continued, suggesting the target turn is a non-alignment.

One might ask why this is a non-alignment, as the understanding check is necessary for B to be sure of having understood C. In order to argue for the non-aligning nature of this understanding check, it has to be taken into account that this analysis is based on the fine grained turn-by-turn analysis. From a more global perspective of alignment, the understanding check is used as a service of a longer-term project of "aligning" with the other speaker. These two ways of aligning have to be distinguished. Here, the focus is on the turn-by-turn structure of alignment. In this sequence, Participant C orients to the short turn by Participant B and treats it as an understanding check that needs immediate clarification. Therefore, the prior agenda is abandoned for a short time and Participant C confirms the understanding of Participant B that he talks about a specific topic. The analysis of this sequence suggests that an understanding check can be analysed as a non-alignment on a turn-by-turn basis.

Looking at the gestures from Participant C, one can see that the gestural agenda is not continued either. Note that Participant C's nod gesture at the very end of the word "parallelogram", uttered by Participant B (continued after Participant B's turn) is different to the hand gesticulation gesture at its beginning while Participant B and C speak in overlap. This gestural non-continuation also shows orientation to the target turn as non-alignment.

4.1.7 Collaborative completion

The examples above of overlapping target turns illustrate lexical shapes other than tokens such as "uh huh", "mm-hmm", "yeah", "right", "okay", "perfect", etc., of which the first two ("uh-huh" and "mm-hmm") can be seen as the prototypical continuers (as described by Schegloff (1982) and by Gardner (2001). There are longer first components of turns such as "oh really" (Extract 8) and "just the parallelogram" (Extract 9). Each is followed by a specific orientation of the prior speaker towards the target turn. But, as shown in Extract 10, it is not always the case that longer turns elicit a specific orientation. Participant A produces an extended turn and comes to a point where he struggles to complete a part of the turn ("... but it's not (.) ..."). He recycles parts of it in "you know it's not". Here, Participant B comes in with "it's not too bad", the target turn, which could be a sensible continuation of the prior speaker's turn. At the same time (in overlap) Participant A continues with his own version of a possible continuation. He continues even beyond the turn of Participant B, who does not challenge this continuation. Such anticipatory or "collaborative

completion" (Lerner, 2002) of turns seems to be a sequence that allows the prior speaker to continue on the same agenda after that turn, i.e. is an alignment.

Extract 10: "slightly bigger" (B-07:37)

[1]

Gest_A	*body movement*
A	butum °hso they uhyeah i think the mouse is slightly bigger but it's not (.) you know
Gest_B	*nod and blink*
Gest_C	*nod*
	(0.26)

[2]

Gest_A	*nod and blink*
A	it's not it's not hugely bigger s o°h
Gest_B	*nod and blink* *nod and blink* *nod* *nod and blink* *blink*
B	it's not too bad uh-huh i c* i can i can see where that's
Gest_C	*nod*
	(0.08) (0.32) (0.14)

[3]

Gest_A	*nod* *frown shrug*
A	yeah and if it ever gets to be a problem in the pilot ...
Gest_B	
B	useful °hh no

Considering the "uh-huh" from Participant B (line 2) as target turn, it can be interpreted as a case similar to the example discussed above (cf. Extract 4), where the second speaker (here B) leaves space for the prior speaker to add more on the prior agenda, before starting a new agenda herself.

A case of something similar to a collaborative completion is taken from Extract 11. Before analysing this extract according to alignment and non-alignment it has to be explained what the target IP is, as the whole utterance of Participant B "they are the paid yeah and" can be divided into two parts with respect to the verbal, the prosodic (two IPs) and the non-verbal content. Participant B overlaps Participant A's end of his extended turn with "they are the paid" which is accompanied by nodding, while the following "yeah and" is accompanied by head shake and blink. This suggests that the verbal content can also be split in two or even three, where at least "they are the paid" can be separated from the rest. The target IP is here the first part: "they are the paid".

Participant B recycles parts of the prior speaker's talk "they're like" in her "they're the paid" (line 2). She is correct in predicting that what Participant A is projecting with the choice of the word "obviously" is related to money ("a few hundred pounds"). Participant B immediately adds "yeah and" while Participant A takes an inbreath and starts to continue on his agenda, that is searching for software that is free and not costly. This suggests that the target turn, the attempt of producing a collaborative completion, can be analysed as an alignment. The point of this example is to support the analysis of Extract 10 with an additional example. Of course, it is not the same, but similar enough to support the same analysis: a collaborative completion or choral co-production that is treated as alignment.

Extract 11: "obviously" (B-11:05)

[1]

Gest_A	*nod* *shrug*
A	um but i kn* i know adobe and things have faculties that do it but obviously they're
B	what kinds of

[2]

Gest_A	*nod*						*shrug*		*nod*
A	like a few	hundred poun	d sh	h °u	m so	°h oth* but	other i	i'd b*	i'd be very surpri
Gest_B		*nod*	*head shake*	*blink*	*head shake*				
B		they are the pai	d ye	a hand					
Gest_C				*nod*					

[3]

Gest_A		*raised eyebrows*			*nod*
A	sed if there wasn't a free package for doing it	(.) u:m		i ha* i have a	feeling ..
Gest_B			*nod and blink*		
B			uh-huh		
		(0.31)(0.55)			

The last two examples are very close to what Lerner (2002) calls “choral co-productions”. Although the second speaker does not produce the same utterance as the first speaker and the start of the second speaker’s utterance is slightly off the start of the first speaker’s utterance, the design of that sequence is similar to Lerner’s observations, where a second speaker’s formulation is a co-production with the first speaker who continues on the prior agenda, e.g. a list (cf. Lerner, 2002). In both extracts above, it is treated by the prior speaker as a permission to continue on the prior agenda. In Extract 10, Participant A finishes the overlapped turn till to a complete whole with a final inbreath “`it’s not hugely bigger so °h`”. In Extract 11 the same structure (completion of a turn with inbreath) can be observed, supplemented with “`um so °h`” and further talk on the prior agenda.

4.1.8 Second component of a turn

We have seen many examples where the second speaker produced two consecutive turns. The first was either followed by continuation on the agenda from the prior speaker or a stretch of silence developed. The second was used to initiate a new agenda. In the following examples, Participant B is the second speaker:

From Extract 4: “software” (B-03:34):

```
A:    ... for that part (0.38)
B:    mm-hmm
A:    um
      (0.55)
B:    so I’ve got a new ...
```

From Extract 5: “reasonable” (B-07:53):

```
C:    ... for the tracking purposes °hh hh°
B:    yeah
      (0.38)
B:    although if if you had ...
```

From Extract 6: “moved over” (B-00:03):

```
C:    ... the other eye link s[o uh
B:                            [yeah
      (0.54)
B:    I mean
C:    ...
```

From Extract 10: “slightly bigger” (B-07:37):

```
C:    ... it's not hugely bigger so °h
B:    uh-huh
      (0.46)
B:    I c* I can I can see ...
```

In these examples, the first component and the second component are separated either by a verbal utterance from the prior speaker or by a stretch of silence.

One more such example is contained in Extract 12 (line 2). The topic of that extract is the search for a certain type of software (a "browser" that can edit SVG files). Prior to the target turn of interest, Participant A explains that he has "a feeling" about one specific browser "that can actually edit SVG files". The first IP "okay" of Participant B overlaps the last bit of the prior speaker's turn and is followed by a short stretch of silence (0.21) and a second IP "so i wouldn't do any". The second one is undoubtedly initiating a new action. It starts a clear recommendation on how Participant A should proceed with the search for a browser (by not doing any work to find it).

Extract 12: "browser" (B-11:19)

[1]

Gest_A		*nod*	*nod*	
A	the((w_)) three((c_))	browser	does it h° have a feeling that	that can actually edit
Gest_B		*nod and blink*	*nod and blink*	
B		yep mm-hmm		

[2]

Gest_A			*frown*	*nod*	
A	((s_v_g_)) files				
Gest_B		*blink*			
B		okay	so i wouldn't do any	any work to find out	what'll output ((s_v_g_))
		(0.21)			

But sometimes a first component of a turn is immediately followed by a second component without a temporal break in between. Essentially, the turn of the second speaker starts with one IP and is followed directly by another IP. The question which of these two IPs would be responsible for the initiation of a new agenda is difficult to answer, as demonstrated by the example in line 1 of Extract 13. Participant A comes to a transition relevance place, followed by silence and a short turn ("mm-hmm") of Participant B. Participant C overlaps that turn with the first IP "uh yes" and launches directly into the next IP "with the um". The first IP could be one of the first components which we have seen in the previous examples. But from the transcript alone this cannot be proven, as the prior speaker does not continue speaking (on agenda or off agenda), nor does a pause develop after the "yes".

Extract 13: "slightly easier" (B-05:31)

[1]

Gest_A		*raised eyebrows*		*blink*		
A	that should make that (.)	slightly eas		ier		
Gest_B					*nod and blink*	
B						mm-hmm
C						uh yes with the um (.)
				(0.11)	(0.42)	(0.22)

[2]

C	the bleep was a hardware bleep originally (.) ...

4.1.9 Laughter

It is possible that the second turn provokes successive laughter, as in Extract 14. Participant C found a previously unidentified source of a bleep sound (which could have been a technical artefact).

Extract 14: "oops" (B-05:35)

[1]

C	the bleep was a hardware bleep originally (.) and so it comes from the ((p_c_)) which is

[2]

A				((sound: laugh))			
Gest_B			*raised eyebrows*				
B			oops	((sound: laugh))		well	you found that then
Gest_C		*reised eyebrows*		*body movement*			
C	outside the sound-proof b o x hh°			((sound: laugh))			
				(1.56)	(1.67)	(0.21)	(1.17)

There is something curious about it and reporting the situation, Participant C comes to the end of the turn with "`... which was outside the soundproof box`". This could be regarded as the punchline, or it could be a point where some other recognition of this curious finding could happen. But laughter is not provoked immediately. Only after the "`oops`" from Participant B, all three participants burst into laughter. Therefore, this "`oops`" is the turn towards which the successive laughter orients, suggesting that the "`oops`" is treated as non-alignment.

An example of a deviant case in Sikveland's (2012) study has also involved laughter at successive positions from the speaker who was at first position. In that study, laughter treated the prior turn "as providing a sarcastic stance" (p.92). Therefore laughter shows special treatment of the prior turn that can therefore be classified as non-aligning.

4.1.10 Summary

In summary, the observations above show that the prior speaker has a tendency after a turn of a second speaker to either continue on the prior agenda or to orient towards the second speaker's turn. If none of these possibilities are chosen by the first speaker, a lapse (Sacks, Schegloff, & Jefferson, 1974) may develop, after which any speaker can take the floor, or the second speaker starts immediately a new agenda (with a new intonational phrase).

These observations are also in line with findings of other researchers who support the hypothesis that there are resources that can be "employed for the fundamental conversational activity of displaying the sequential status of an immediately next turn either as a continuation of prior talk, or as the beginning of a new project." (Szczepek Reed, 2012a, p.14). Szczepek Reed calls this fundamental conversational activity a "generic social action format" which is "irrespective of the *precise* prosodic pattern, irrespective of the specific action being accomplished, and irrespective of sociolinguistic factors, such as variety, gender and age" (p.16).

The treatment of the second speaker's turn by the co-participants directly after that turn gives a clue how it has been interpreted by the participants. It is possible to divide the treatment of the second speaker's turn into two main classes. In one class, the prior speaker treats it as an invitation to continue on the prior agenda. In the other class, the prior speaker treats it as an initiation of a new agenda, which requires special orientation.

This partly answers research question (RQ1), asking what the sequential correlates of aligning and non-aligning actions are. They occur in adjacent turn sequences, located at transitions from one speaker to another and they are organised according to the context of overlap, silences, or

further talk from the same speaker. However, the relevance of these actions to the interactional participants and the development of the interaction has not been attested yet.

4.2 Classification of agenda aligning vs. agenda non-aligning actions

The previous section demonstrated the procedure used to collect adjacent turn pairs which have the potential of exposing aligning or non-aligning actions. It was shown that the orientation of participants can help to interpret the interactional context of adjacent turns in order to make predictions about the social action performed in these sequences. The observations demonstrate that patterns exist in the organisation of talk and gesture. These patterns will be discussed in more detail here.

The basic difference between alignments and non-alignments is the continuation – or non-continuation – of the prior speaker's agenda after the target turn. In this section we seek to establish a set of interdependent criteria which can be used to distinguish aligning from non-aligning second turns. These can then be used by independent annotators as operational criteria when classifying turns as aligning or non-aligning.

One requirement of this classification exercise is that the annotators should not have access to prosodic information. This is because prosodic features will be dependent variables in the study of prosodic matching and alignment in Chapter 5. The annotators will therefore not listen to audio files of the interaction extracts to be annotated. Instead they are asked to base their decision on the orthographic transcripts alone.

In the following, the sequential correlates, or the sequential consequences for the interaction of the social actions of alignments and non-alignments are presented, i.e. the differences in the sequential organisation of the two categories are developed. This requires closely analysing the environment of adjacent turns, in order to establish the criteria which make a second turn an alignment and which make a second turn a non-alignment. It should be possible for other annotators to take these criteria and arrive at the same categorisation. The broad classification into alignment and non-alignment, warranted from a detailed interactional analysis, does not exhaust the interactional potential of second turns, which may simultaneously perform other actions, as will be seen.

The prior speaker's action trajectory

In analogy to what Szczepek Reed calls "action trajectory" (Szczepek Reed, 2012a, p. 13), the term "agenda" will be used in the following. As reported in Chapter 2 Section 2.1.2, there is a general consensus in the literature that aligning actions are treated by the other participants – especially by the prior speaker – as an opportunity to continue on the agenda of the speaker who had the floor prior to that alignment. Non-aligning actions are treated by the other participants as a turn that needs special treatment. In the latter case, the prior speaker does not have the opportunity to continue on the prior agenda, but orients towards the non-aligning turn in a special way. A non-aligning action is also attested when the prior speaker is not given the opportunity to continue speaking, for example by a floor challenging move, i.e. when the second speaker initiates a new agenda with the adjacent turn itself.

It is however not a straight forward task to identify whether the prior speaker continues after the second turn on his agenda or whether the second turn is subsequently treated in another specific way. Several questions arise: How can we determine if some talk is on the agenda of some prior talk? How can we determine if some talk treats some other talk in a specific way? Does it simply depend on the lexical choice? Or does it also depend on the sequentially surrounding context? Is it possible that not only the subsequent talk plays a major role, but that also other structural elements have an influence, such as inbreaths, outbreaths, stretches of silence, gestures, developments and resolutions of overlap – and their sequential interplay?

With sequential interplay we mean the relatedness of such elements. For example, it is very likely that the verbal content of the prior speaker's talk after the second turn will be related with the verbal content of the prior speaker before or during the second turn. Similarly, it is very likely that the gesture of the prior speaker after the second turn will be related to the gesture of the prior speaker before or during the second turn.

The necessary structural elements, as presented above (Section 4.1), will be discussed in more detail as follows. We begin with the verbal talk after the second turn.

4.2.1 Verbal talk after the second turn

The extracts in the previous section (4.1) have illustrated some different sequential organisations of adjacent turns during conversation. One of the environments was talk from one speaker, followed (or not) by a pause, breathing, hesitations; then a second turn, followed (or not) by a pause and more talk from the prior speaker, which we call "successive talk".

4.2.1.1 Verbal continuation of prior talk in successive talk

The primary goal in determining whether the prior speaker continues after the second turn on the prior agenda or not, is to find evidence for it in the successive talk of the prior speaker. If that talk is on the prior agenda, the second turn can be classified as alignment. If the successive talk orients towards the second turn in some specific way, the second turn can be classified as non-alignment.

In Extract 15, Participant A says: `"i'd be very surprised if there wasn't a free package for doing it"` followed by a short silence and hesitation. The second turn `"uh-huh"` (end of line 1) from Participant B is followed by more talk from A, who starts with a truncated `"i ha*"` and restarts `"i have a feeling that the:"` which is here the successive talk. This talk can be said to be on the prior agenda, as Participant A talks about `"i have a feeling"` which is topically related to the conditional `"i'd be very surprised"` from the prior talk. Participant A also mentions a browser from the W3C organisation which complies with the requirement of a `"free package"`. The successive talk does not orient to the `"uh-huh"` in any other specific way.

Extract 15: "free package" (B-11:12)

[1]

Gest_A	*nod* … *raised eyebrows*
A	i'd be very surpri sed if there wasn't a free package for doing it (.) u:m
Gest_B	*nod and blink*
B	uh-huh
	(0.31)(0.55)

[2]

Gest_A	*nod* … *frown* … *nod* … *nod*
A	i ha* i have a feeling that the: um amaya the((w_)) three((c_)) browser does it h° have
Gest_B	*nod and blink* *nod and blink*
B	yep mm-hmm

[3]

Gest_A	
A	a feeling that that can actually edit ((s_v_g_)) files
Gest_B	
B	okay
	(0.21)

The turn “yep mm-hmm” of Participant B in line 2 is followed – and even partly overlapped – by more talk from Participant A. Again, the “have a feeling” is recycling “have a feeling” from the prior talk. It is the same topic. That indicates that the successive talk is a continuation of the prior speaker’s agenda.

At the beginning of Extract 16 we find another example where the successive talk of the prior speaker after the second turn is on the prior speaker’s agenda. Participant C has “sorted out travel et cetera to munich”. He starts to list for whom the travel has been arranged, when Participant B says “oh yeah”. The successive turn “for ellen (.) john and myself” (at the end of line 1, beginning of line 2) does no more than finish the previously started turn “for °h e*”, which C interrupted, i.e. to specify the persons for which the travel has been sorted out.

Evidence for agenda continuation is the “for” of Participant C prior to the “oh yeah” from Participant B, by its recycling in “e* for ellen ...” in the successive talk from Participant C.

This indicates that the second turn “oh yeah” was not treated in any other specific way than letting the prior speaker continue on the prior agenda. In this case it was not necessary to look far from the target turn in order to make the decision about alignment and non-alignment. The action trajectory “travel arrangements”, followed by the mentioning of the “travel participants” can be identified just before the target turn. The short interruption of the turn “for °h e*” and its immediate continuation in the turn “for ellen john and myself” is sufficient to identify the continuation of this action trajectory.

Extract 16: “travel et cetera” (B-18:34)

[1]

Gest_B	*nod and blink nod*
B	oh yeah
Gest_C	*nod*
C	i've sorted out (0.63) u h (0.49) travel et cetera to munich hh° (0.39) for °h e* for ellen (.)

[2]

Gest_B	
Gest_C	*raised eyebrows*
C	john and myself (-) we're all off on the (0.61) fourteenth (0.09) (0.33) of december (0.83)

[3]

Gest_B	*nod* *nod and blink* *nod and blink*
B	mm-hmm and john's going mm-hmm oh that's good
Gest_C	*nod* *nod*
C	coming back on the fri day the sixteenth yes (0.24) (0.33) (0.47) (0.92)

[4]

Gest_B	*nod*
B	he'll be useful for the ((w_ p_)) seven stuff so (1.06)
Gest_C	

4.2.1.2 Verbal non-continuation

Different to the verbal continuations above is the verbal non-continuation that can be observed in line 3 in Extract 16. The target turn is “and john’s going” that is followed by C’s successive talk “yes”. In line 1 and 2, C has already specified for whom some travel

arrangements have been sorted out, including "john". This may have passed Participant B's attention for some reason. After C says the dates for when "we're all off", Participant B who may have forgotten who these "all" were, apparently asks for confirmation "and john's going", overlapping C's elaboration of when they are all "coming back", namely "on the friday the sixteenth". The "yes" from C is syntactically detached from the prior talk "coming back on the friday the sixteenth". The two turn components from C seen together ("coming back on the friday the sixteenth yes") does not make sense, i.e. the two components can't be interpreted as two parts of the same pragmatic action. Only if the "yes" is not considered as a specific orientation to Participant B's overlapping turn "and john's going", it becomes clear that the "yes" is not a continuation of the preceding turn component of C. It treats B's turn as an initiation of something that requires Participant C to abandon his activity in progress and to orient to it in a specific way. The clearly articulated "yes" – note: not "yeah", "yep" or the like – can be identified as a positive clarification. This suggests that interjections, such as the target turn "and John's going" can be analysed as non-alignment.

Again, long-term activity, i.e. the giving and receiving of the report, is overshadowed by B's short-term project which is to understand the travel arrangements, i.e. interrupting for clarification on a turn-by-turn basis. This gives rise to some adjacent turns where there is local non-laignment (as here). The overall understanding follows, which is demonstrated with an immediate and overlapping "mm-hmm" from Participant B, accompanied by nodding and blinking and followed by a strong positive assessment "oh that's good" – treated as such by continued nodding by C and a long stretch of silence that develops.

Considering the turn "mm-hmm" as target (beginning of line 3 after Participant C's "we're all off on the (0.6) fourteenth (0.4) of december (0.8)"). Participant B's "mm-hmm" overlaps with Participant C's "coming back", which extends into the successive talk "on the friday the sixteenth". Both parts the continuation (in overlap with the target turn) and the subsequent talk build one unit, which is syntactically continuous and coherent with the prior agenda (describing the travel times for outbound travel and return). This suggests that the "mm-hmm" is treated as an alignment.

In Extract 17, Participant A comes to a transition relevance place (line 2). The sentence "... next time i'm working on this uh GDF stuff" can count as syntactically complete. Participant A adds a hesitation "um" just before the "yeah" from Participant B. The successive talk from Participant A starts with "so". In the following talk, Participant A expands on what "working on GDF stuff" contains. Something similar to syntactic continuation can be observed here, when the prior speaker starts the subsequent talk with a backwards connecting token such as "and", "but", "or", "so", etc. The "so" in this case initialises the continuation on the prior agenda.

Extract 17: "GDF stuff" (B-19:23)

[1]

Gest_A	*body movement* *shrug*
A	if there's any problems gonna tr* gonna try and fix them °h if there isn't any problems next

[2]

Gest_A	*body movement*
A	time i'm working on this uh ((g_d_f_)) stuff u m s o so i've i've i pu* uh put
Gest_B	*nod and blink*
B	yeah
	(0.28)

[3]

Gest_A	*raised eyebrows*
A	on a ((d_t_d_)) and a sample ((x_m_l_)) file on the wiki for the ((g_d_f_)) °h ...

There is dysfluency at the beginning of the subsequent talk "`so i've i've i pu* uh put`". It could be argued that a turn exposing these phenomena is in itself an orientation to the prior turn. For example it could be an acknowledgment that this turn was potentially a bid for the floor, i.e. it could have turned into one. However, it remained short and non-competitive (Kurtić, Brown & Wells, 2009), leaving the floor to the current speaker.

There are a number of other instances where the subsequent talk (also from other participants) starts with hesitations or truncated words with lexical repeats which are similar to the phenomenon of dysfluency.

Line 2 of Extract 9 is an example of dysfluency from a non-aligning sequence. Participant C says "`it was f*`" being overlapped by B's "`just the parallelogram`". Subsequently this truncated "`f*`" is not restarted and not finished.

Line 1 of Extract 16 is an example from an aligning sequence. Participant C says "`for °h e* for ellen john and myself`" with a truncated "`e*`" that is restarted and finished to "`ellen`" after the overlapped part.

Line 2 of Extract 15 is an example of dysfluency from an aligning sequence. Participant A says "`i ha* i have a feeling that ...`" after Participant B has produced the aligning "`uh-huh`" (line 1). The two examples suggest that the continuation of the prior activity in progress is sometimes accompanied, or started with dysfluency, without special orientation to the dysfluency.

There are also cases where the successive speaker is the target speaker herself and starts with dysfluency.

Line 2 of Extract 10 is such an example. Participant B says "`i c* i can i can see where that's useful`" after her own aligning "`uh-huh`" that resulted in a stretch of silence. An attempt to account for this dysfluency of a truncated "`i c*`" that is restarted "`i can`" and repeated in a second "`i can`", would be that the prior continuer "`uh-huh`" did not achieve a direct verbal continuation of the prior speaker and that the turn fell back to Participant B. She was not directly prepared to speak, resulting in a mix of dysfluency (truncation and stuttering).

Line 2 of Extract 5 is a similar example. Participant B says "`although if if you had ...`" after the previous aligning "`yeah`" did not achieve a direct verbal continuation of the prior speaker. Again, the repetition of the "`if`" might indicate that Participant B had not properly finished planning what to say and was not entirely prepared to take the speaking turn.

A further example of dysfluency can be observed in Extract 18. After the "`mm-hmm`" of Participant B and a short "`so`" of the prior speaker, a relatively long stretch of silence (1.6s) evolves after which Participant B starts with a mix of "`um`", silences, "`yeah`" and other unintelligible vocalisations. The "`so`" from Participant A has many characteristics in common with a so-called "trail-off so" (Local and Kelly (1986); Walker (2012).)

Extract 18: "double check" (B-15:11)

[1]

Gest_A	*shrug*		*shrug*		*nod*			
A	i can easy	double-check	if	uh nijmegen can't find anything	°h	s o		
Gest_B					*nod*			
B					mm-hmm			um (-)
							(1.6)	

[2]

Gest_A		*frown*	*frown*	
Gest_B				*frown*
B	yeah (.) (uh g) i haven't	looked through other mail	i did see something	from marloes

Initially it seems as if the prior speaker (Participant A) continues with the "so", but because nothing else follows, a long stretch of silence (1.6) develops. As Participant A does not add any further talk, the floor falls back automatically to Participant B. Hesitations, silences and truncated word beginnings are the consequence. This indicates that Participant B was not prepared to take the floor at that stage and that the "mm-hmm" was designed to let the prior speaker continue instead.

In Extract 19 (reproduced from Extract 10), Participant A comes to a transition relevance place with "it's not hugely bigger" and the "so" is directly attached to the end of it with a following inbreath. Again, a pause develops after the "uh-huh" from Participant B. This pause (0.46) is shorter than the one in the previous example after the "mm-hmm" (1.6). The subsequent speaker is again the same (the "uh-huh" speaker) and starting in a dysfluent way ("i c* i can i can").

Extract 19: reproduced from Extract 10: "slightly bigger" (B-07:37)

[1]

Gest_A				*body movement*
A		butum °hso they uh		yeah i think the mouse is slightly bigger but it's not (.) you know
Gest_B	*nod and blink*			
Gest_C			*nod*	
	(0.26)			

[2]

Gest_A			*nod and blink*						
A	it's not it's not hugely		bigger s o°h						
Gest_B		*nod and blink*	*nod and blink*		*nod*			*nod and blink*	*blink*
B		it's not too bad			uh-huh			i c* i can	i can see where that's
Gest_C				*nod*					
				(0.08)		(0.32)	(0.14)		

Dysfluency however is not a phenomenon that only occurs after aligning second turns where the right to speak fell directly back, as the prior speaker let lapse the possibility to continue speaking. Non-aligning second turns too can be followed by successive talk that exhibits hesitations or truncated words, e.g. from the prior speaker. Extract 20 illustrates such a case. After speaker B's turn "what documentation", the prior speaker takes an inbreath and starts with "some b* uh" and continues with "JAST report". Two explanations for the dysfluency can be suggested: Either, the "some" is partly recycling the "something" from the prior talk in a truncated form. The transcribed "b*" would then represent an audible release of the bilabial closure [m] of the truncated "some". This would indicate that Participant C abruptly interrupts the envisaged continuation, demonstrating that after some time of thinking (0.79s and the word "some") C turns his attention to B's "what documentation" in order to provide a response as adequate second pair part to a question. Or the "some b*" is already related to the following "JAST report". The latter would suggest that Participant C was not prepared for giving a response to B's "what documentation" and confused the starting letter of the report ([b] vs. [ʤ] for "JAST"). Both indicate that the successive turn does not continue the prior speaker's agenda. A"JAST report" is a possible response, and treats speaker B's turn "what documentation" as a question. The successive turn orients to the second speaker's turn in a specific way and does not continue the prior agenda.

Further support for the classification of that turn as non-alignment is that it is not possible to connect the syntax of the first turn of the prior speaker with his successive turn. The version"... on this project or something °hh some b* uh JAST report" does not make sense. Even if the "some" is seen as a lexical repeat of the "something", there is a logic break between "on this project or something" and "JAST report".

Extract 20: "five percent" (B-36:38)

[1]

Gest_C	*body movement*	*body movement*
C	you're down for five percent of your fifteen months	on this project or
	(0.06) (0.29) (0.1)	

[2]

Gest_B		*blink*			*body movement*
B		what documentation			o hok ay
Gest_C		*body movement*	*head shake*	*blink*	
C	something		°hh some b* uh jast report		
	(0.42) (0.26)		(0.79)		

In Extract 21, Participant C starts the subsequent talk in a similar way: inbreath, some truncated talk and hesitations ("°h uh the i'll (.) mm"). The "not sure" is syntactically disjoint from what comes before. It neither fits with "uh the i'll" nor with "yes i think so". It treats the "where are they" from Participant B as a question that makes a response relevant, even though the response comes relatively late.

Extract 21: "folk wisdom" (B-32:12)

[1]

Gest_B		*frown*				*nod*
B		where are they				okay
Gest_C	*nod*			*head shake*		*body movement*
C	((sound: click)) yes i think so		°h uh the i'll		mm not sure °h	h
			(0.61)	(0.11)		(0.52)

[2]

B	well (.) they might have the folk wisdom o n			the complexity ...
Gest_C	*nod*	*nod*	*nod*	
C	hhh°			
	(1.03)		(0.46)	

These two examples have in common that the second turn starts with a WH-question word ("what" and "where"). It could be argued that these specific words make the turns questions, which require specific treatment and that the analysis of the context is superfluous. But a similar delayed uptake and orientation to an utterance can also be observed after second turns without question word or question syntax, as the next example demonstrates.

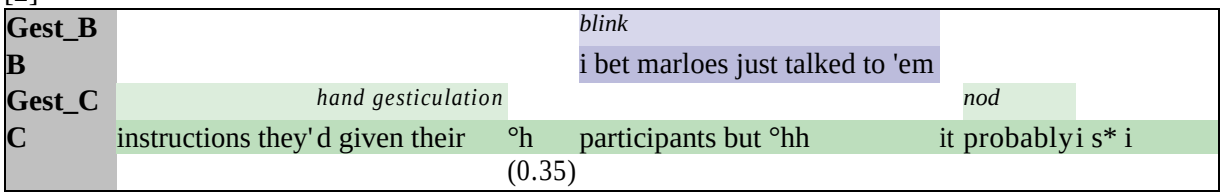
In Extract 22, line 2, the successive speaker starts an utterance, which is then truncated for subsequent talk that orients toward the second turn with "it probably i s* i suspect so". The beginning of the subsequent talk "it probably" is syntactically in line with the prior speaker's talk and its continuation (in overlap with the second turn). But the following restart "i s* i suspect so" introduces an immediate syntactic break and a break in the speaker's agenda. The second part of that successive talk then orients towards the second turn and treats it in a specific way, i.e. it does not continue on the prior speaker's agenda. This indicates that not only "questions" require special treatment, but sometimes also statements do.

Extract 22: "just talked" (B-25:33)

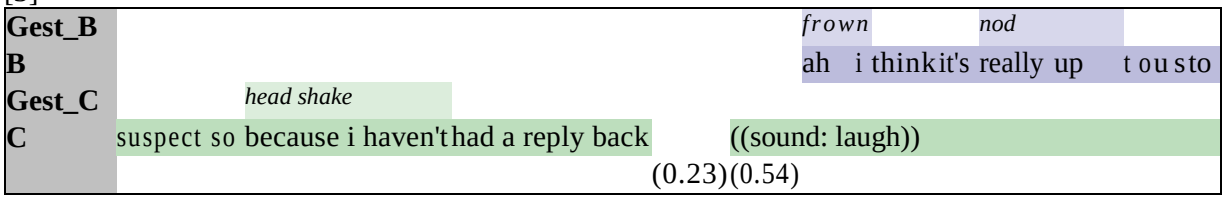
[1]

Gest_C	*body movement*	*frown*	*head shake*		*head shake*
C	actually i i did e mail (-) nijmegen to ask if they'd had a set of				
			(0.46)	(0.14)	(0.11)

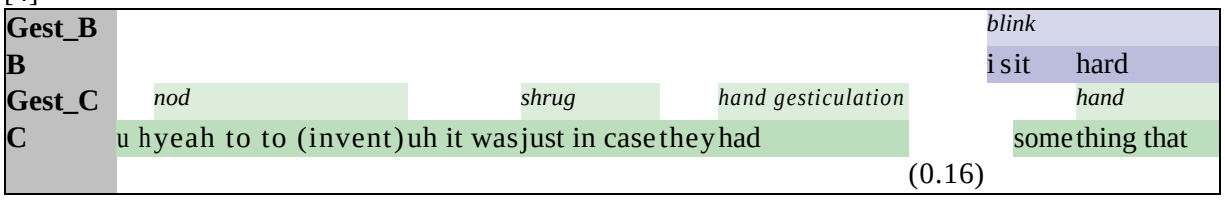
[2]



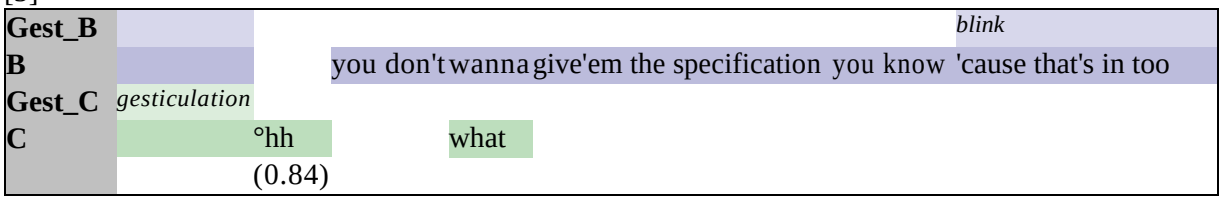
[3]



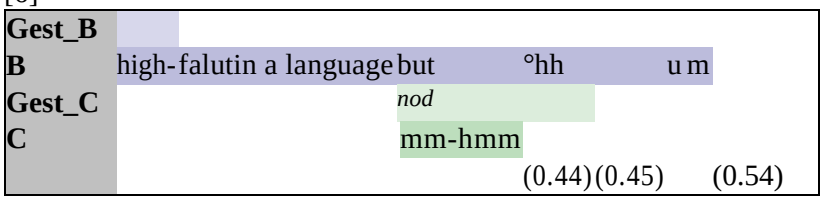
[4]



[5]



[6]



The second turn of Participant B in this extract "`ah i think it's really up to us to`" is followed by the prior speaker's "`uh yeah`", indicating that the prior speaker in the successive talk orients towards the second speaker's turn in a specific way. This suggests that the target turn is a non-alignment.

One further turn by Participant B "`is it hard`" (line 4) is produced with question syntax. It receives special treatment by the prior speaker with a return-question "`what`" (line 5).

4.2.2 The prior speaker's gestures

An important issue to consider is whether head gestures are inherently different from manual ones. On the one hand, they seem more permeable and more extendable than gestures performed with the hands. On the other hand, the flexibility of the hands with the various configurations a hand can form, the gesticulation of a hand is potentially much richer. Therefore it was decided to analyse one of the most fundamental properties shared by all gestures: its continuation or non-continuation. The range of gestures considered was restricted to basic head gestures, hand gesticulation, body movement and facial expressions (see Section 3.1.3).

In Section 4.1 it was observed that a gesture continuation sometimes coincides with a verbal continuation of the prior speaker's agenda:

In Extract 15: “free package” (B-11:12) (line 2), the “`yep mm-hmm`” of the second speaker is simultaneous with a nod of the prior speaker, and it is also followed by a nod from the prior speaker afterwards.

In Extract 8: “eye” (B-07:09), the “`oh okay`” is simultaneous with body movement of the prior speaker, which carries over into the pause and the start of the third speaker.

In Extract 4: “software” (B-03:34), the “`mm-hmm`” is preceded, accompanied and followed by a nod from the prior speaker.

In all these examples, instances of simultaneous gesture and verbal continuations of the prior speaker’s verbal or gestural agenda have been identified. Following these observations, it may be the case that gestural continuation or non-continuation reflects the treatment of the second turn as alignment or non-alignment, even without verbal evidence for it. This is explored further in the following section.

4.2.2.1 Gesture continuation

Extract 23 gives an example of such sequential organisation where no verbal talk, but a gesture of the prior speaker carries over into the gap after the target turn “`yeah`” (line 1). Participant C produces the turn “`and it’s all worked`” with body movement during the last two words (line 1). Participant B’s “`yeah`” overlaps parts of it and she blinks. Participant B takes over the floor with verbal hesitation / dysfluency. The body movement of Participant C continues half a second into the following silence. It suggests that the target speaker (B) left room for the prior speaker to continue. Here the prior speaker continues the gesture, suggesting a continuation of the gestural agenda. This further suggests that the “`yeah`” is treated as alignment.

Additionally to the aligning “`yeah`” that allowed C’s continuation, which was done gesturally but not verbally, the successive turn by B starting with a stretched “`s::o::`” can be analysed as an orientation to a possibly incomplete prior. C’s “`and it’s all worked`” is questioned by B’s “`so did he get his data in or is this put his schedule back`”. This indicates that more information could have been expected from the prior speaker after the alignment “`yeah`”.

Extract 23: “experiment” (B-02:31)

[1]

Gest_A		*nod*		
A			yep	
Gest_B		*blink*		*blink*
B		yeah	s:: o:: uhi* uh did he get his	
Gest_C	*body movement*	*body movement*		
C	so he's run an experiment (0.32) and it's all worked			
			(0.48)(0.34)	

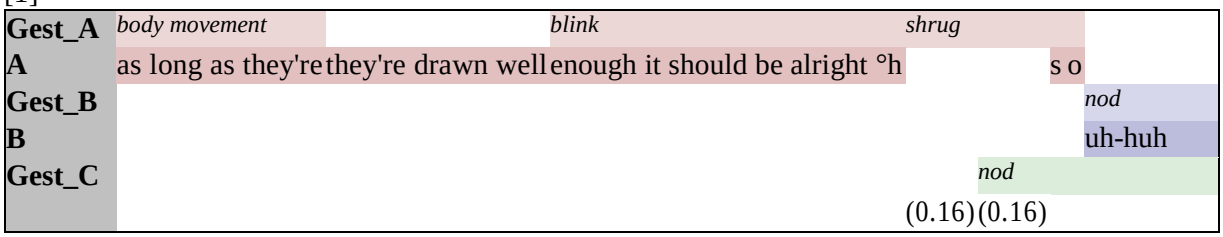
[2]

Gest_B			
B	data in or is this (0.21) put his schedule back		
Gest_C			*head shake*
C		((sound: click)) uh he's he's done ...	
	(0.12)		

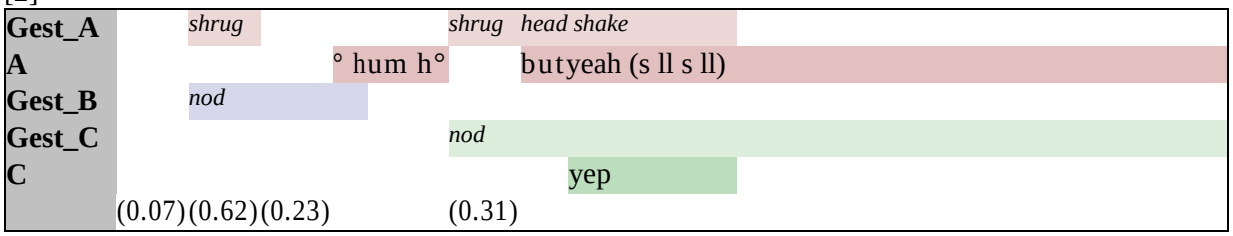
Not only body movements and nods are continued but also shrugs, as of Participant A in Extract 24. Participant A comes to a transition relevance place with “`it should be alright`”, followed by a shrug and “`so`”. The “`uh-huh`” from Participant B is followed by another shrug from the prior speaker. Synchronous to the second shrug from Participant A, Participant B performs a head nod, followed by A’s hesitation “`°h um h°`” and another shrug.

Extract 24: "drawn well enough" (B-13:58)

[1]



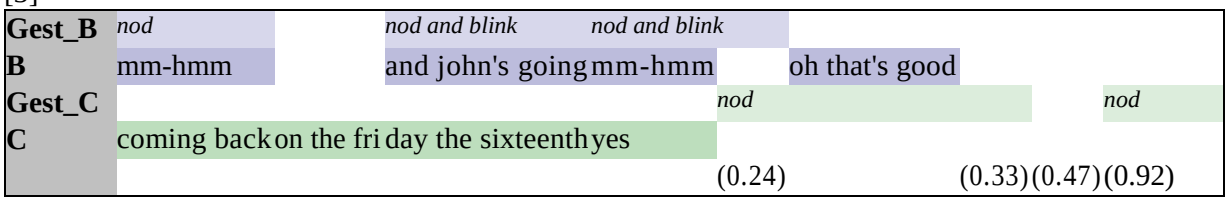
[2]



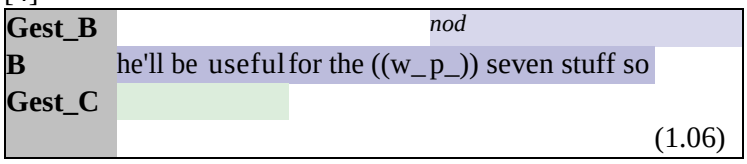
Participant A's verbal continuation with hesitation and the gestural continuation of repeated shrugs indicate A's treatment of the "uh-huh" as allowing to continue on the prior agenda. Participant B does not start a new agenda at this point, nor does Participant A treat the "uh-huh" in any other specific way. The interpretation that the repeated shrug is a sufficient continuation is supported by the developing pause and the prior speaker's inbreath, hesitation, outbreath and another shrug, indicating that the right to speak or to gesture returned to A and suggesting that the "uh-huh" is an alignment. These continuations of the prior gesture, together with the hesitations and in- and outbreaths can be analysed as an orientation to the lack of uptake by another Participant. This is even more evidence that the "uh-huh" is designed as a turn that allows continuation of the prior speaker, i.e. an alignment. Another example comes from Extract 16, reprinted here in Extract 25:

Extract 25: reproduced from Extract 16: "travel et cetera" (B-18:34)

[3]



[4]



Participant C had the prior turn ("yes") and nods afterwards (line 3). The nod continues during Participant B's "oh that's good" and beyond. For a short while (0.47) the nodding is interrupted, but restarts and does not stop until Participant B is already well in a new turn.

In Extract 26, Participant A comes to a transition relevance place with "... is what gets played °hh so" (line 1). The "right" from Participant B (line 1) is preceded, accompanied and followed by head shakes from Participant A. His successive talk "so you're say'n just change sound ..." could be analysed as an orientation to B's turn being an understanding check, if the "you're say'n" referred to Participant B personally. However, here the "you" is a general reference to a person who deals with software and who wants to tell it what sound it should play under certain circumstances. Here,

the software is treated like a person that needs to be told what it should be doing. The overall topic is that a certain sound needs to be picked to be played by the loud speakers in the recording room in order to be able to synchronise the audio channels afterwards. The prior talk is reporting the current recording setup where “whatever the asterisk sound is on windows is what gets played”, followed by “°hh so”, projecting further talk to come.

Extract 26: “asterisk sound” (B-05:54)

[1]

Gest_A	*hand gesticulation* *head shake* *hand gesticulation* *head shake* *head shake*
A	uh whatever the asterisk sound is on windows is what gets played °hh so (uh) s*
Gest_B	*nod and blink*
B	right
	(0.53)

[2]

Gest_A	*shrug* *head shake*
A	so (uh) so you ‘re say’n just change sound to whatever you like just go through the windows
Gest_C	*nod*

[3]

Gest_A	*head shake* *shrug*
A	control panel
Gest_B	*nod and blink* *blink*
B	okay °h h so we want one you can pick up (-) using matlab
Gest_C	*nod*
	(0.2) (0.38)

The expression “just change sound to whatever you like” is a continuation on the prior agenda and does not address the “right” from Participant B in any other specific way. The “so (uh) so you’re say’n” is a lead into expressing how the required sound can be picked: by telling the programme “through the windows control panel”. It recycles the previous “so (uh) s*” which is partly overlapped by B’s “right”. This is further syntactic evidence that the prior and the successive talk from A are continuous, suggesting that the target turn is an alignment.

Further down in the same extract, Participant A comes to a transition relevance place with “... just go through the windows control panel”. Again, the “okay” from Participant B is preceded and followed by head shakes from Participant A. It is only after this sequence that Participant B picks up the floor and starts a new agenda with “so we want one you can ...”.

The previous examples demonstrated that when second turns are treated as alignments, the prior speaker often continues the gesture from the prior turn. This suggests that in parallel to verbal cues, also gestures can give clues about the prior speaker’s agenda and whether that agenda is continued.

Considering how head and body movements affect our understanding of the turn space, the observations above suggest that gesture continuation can be understood as a resource parallel to verbal continuation when treating a second turn as alignment. Even, if the verbal modality is not used and therefore not available for the analysis after the second turn, it is also possible in some cases (see Extract 24) to use the gestural modality alone to as a source of evidence.

4.2.2.2 Gesture non-continuation

The same method also tells us when the gestural agenda is not continued. The four examples from Extract 20: “five percent” (B-36:38), Extract 21: “folk wisdom” (B-32:12) and Extract 22: “just talked” (B-25:33) share this characteristic.

In Extract 27 (reproduced from Extract 20), the turn "what documentation" from Participant B is preceded and accompanied by body movements from Participant C, but the successive gesture is a head shake during "some b* uh".

Extract 27: reproduced from Extract 20: "five percent" (B-36:38)

[1]

Gest_C	*body movement*	*body movement*
C	you're down for five percent of your fifteen months	on this project or
	(0.06) (0.29) (0.1)	

[2]

Gest_B		*blink*			*body movement*
B		what documentation			o hok ay
Gest_C		*body movement*	*head shake*	*blink*	
C	something		°hh some b* uh just report		
	(0.42) (0.26)		(0.79)		

In Extract 28 (reproduced from Extract 21), the prior gesture is a nod during "yes i think so" and the successive gesture is a head shake during "(0.11) mm not sure".

Extract 28: reproduced from Extract 21: "folk wisdom" (B-32:12)

[1]

Gest_B		*frown*				*nod*
B		where are they				okay
Gest_C	*nod*			*head shake*	*body movement*	
C	((sound: click)) yes i think so		°h uh the i'll	mm not sure	°h h	
			(0.61) (0.11)		(0.52)	

[2]

B	well (.) they might have the folk wisdom o n				the complexity ...
Gest_C	*nod*	*nod*	*nod*		
C	hhh°				
	(1.03)		(0.46)		

In Extract 29, (reproduced below from Extract 22), the prior gestures is hand gesticulation during "instructions they'd given their". The successive gesture is a nod during "probably".

Extract 29: reproduced from Extract 22: "just talked" (B-25:33)

[2]

Gest_B		*blink*		
B		i bet marloes just talked to'em		
Gest_C	*hand gesticulation*			*nod*
C	instructions they'd given their °h	participants but °hh		it probably i s* i
	(0.35)			

[3]

Gest_B			*frown*	*nod*
B			ah i think it's really up	t ou sto
Gest_C	*head shake*			
C	suspect so because i haven't had a reply back		((sound: laugh))	
		(0.23) (0.54)		

For all these examples it has been established from the verbal content and sequencing that they are non-aligning. For the gestural content we can state that in all these examples, the prior

gesture is not continued subsequently. Opposed to the gesture continuations we can state that when second turns are treated as non-alignments, the first speaker discontinues the gesture from the prior turn.

We have argued above, that agenda continuation can be identified in two ways. The verbal and the gestural agenda can be continued. However, in some instances the cues from both modalities disagree. For instance, a different image develops when we look at Extract 30. On the one hand, the successive talk from Participant A is on the prior agenda, indicating treatment of the "`uh-huh`" from Participant B as alignment. On the other hand, the successive gestures (nod and later head shake) from Participant A do not continue the prior gesture (body movement), indicating specific orientation towards the "`uh-huh`".

Extract 30: "converter" (B-30:07)

[1]

Gest_A	*body movement*		*body movement*	*body movement*		*body movement*	
A	the jast analyser's the base	the(.)	basis for uh	um ascii to ((g_d_f_))	(0.35)	(0.11)	converter

[2]

Gest_A	*body movement*		*nod*	*head shake*	
A	° hum so what i'm gonna	do is	actually creat	e an output uh a	((g_d_f_)) output part
Gest_B		*nod and blink*	*nod and blink*		
B		uh-huh			

[3]

Gest_A	*raised eyebrows*	
A	to it °h um 'cause	at the moment ...

At first glance, this discrepancy between successive talk (continues on prior agenda) and successive gesture (does not continue on prior agenda) seems to contradict the examples above, where both modalities converge in the prediction of the social action. A closer look reveals that the successive gestures of Participant A (the nod during "`actually`" and the head shake during "`an output uh a`") can also be interpreted as gestures that are simply designed to emphasise the words they accompany. For example Schegloff suggests that head shakes are employed for intensification (Schegloff, 1987, p. 106). Kita (2009) reviews how co-speech gesture can enhance the verbal content (Kita, 2009).

Another example where the predictions of the two modalities are in conflict with each other is found in Extract 31. Participant C continues after B's "`oh that's good`" (line 2) on his own verbal agenda, which is slightly different to the prior agenda, however remaining strongly topically related. This would be indicative for an aligning character of Participant B's turn. On the gestural side, Participant C's gestures after B's turn (nod) and before (body movement) differ, suggesting that the gestural agenda is not continued. This would be indicative for a non-aligning character of Participant B's turn. However, the nod seems to acknowledge Participant B's assessment "`oh that's good`" during the inbreath before continuing on his agenda.

Extract 31: "polygons" (B-08:39)

[1]

Gest_A								*nod*
Gest_B				*raised eyebrows*				
B			oh remind me	oh yes	okay			
Gest_C							*body movement*	
C		((sound: other))	you know there was the	°h yeah	the	collisions	the	polygons
	(0.32)							

[2]

Gest_B		*nod and blink*				
B		oh	that's good			
Gest_C				*nod*		
C	now don't		overlap	°hh	um °h marloes discovered something strange about	
	(0.27)			(0.62)		

The question is how frequent such cases are in the data, where the verbal and visual cues are in conflict. They seem to be an exception to the rule and the frequency with which they occur is relatively low.

According to the examples above of gesture continuation and non-continuation it is suggested that the gestural agenda of the prior speaker seems to be coupled in a similar way as the verbal agenda of the prior speaker. If the prior gesture in the prior talk is continued in the successive gesture we could speak of an agenda continuation in the gestural domain. If the prior gesture in the prior talk is not continued in the successive gesture, the agenda is not continued in the gestural domain. It means that the verbal phenomena are not the only phenomena that can help to determine whether the prior speaker's agenda is continued after the second turn or not.

4.2.3 Summary

First, a collection of turn sequences was made as described in Section 4.1. This collection includes short tokens such as "`uh huh`", "`yeah`" and "`mm hmm`" but also longer stretches, such as "`oh yeah uh huh`", "`uh huh yeah`", "`right uh huh`", "`right okay`", "`oh right yeah you said that`", "`oh really`", "`by`" and "`until you get the`". (See Appendix C for a complete list.)

- Each second turn was analysed together with its surrounding context by reference to the orthographic transcript. The result of this analysis is a collection of the criteria which were used by the author as the basis for the categorisation of second turns into alignments and non-alignments. The main criterion is continuation vs. non-continuation of the prior speaker's agenda. Prior and successive talk could be topically related or not.
- In successive talk, words or stretches of turns could be recycled or not.
- The syntactic organisation could be continuous or disjoint.
- The interactional work of silences, hesitations and false starts, sometimes making an impression of dysfluency, were sensitive to the sequential placement and speaker identity (first vs. second speaker).
- "Turn holding" or "trail-off" conjunctions were used.
- Tokens such as "`so`" and "`but`" were used to connect prior and successive turns.
- The gesture in the prior turn could be continued successively or not.
- Laughter treated the prior turn as the turn with the punchline.

This answers the first research question (RQ1) that asked what the sequential correlates of the social actions alignment and non-alignment are.

In the next section, the numeric results of the categorisation of the whole collection of short turns into alignments and non-alignments are presented. It is also tested if the details above can also be recognised by other annotators when they are asked to make a decision based on the orthographic transcripts (Section 4.3.2).

4.3 Results

4.3.1 Descriptive statistics

In the three AMI meetings, 912 instances of a second speaker's turn after a first speaker's turn were identified. Each was classified as either an alignment or a non-alignment. Table 1 gives an overview.

Table 1: Distribution of alignments and non-alignments according to speakers and meetings.

	Target Speaker										Total
	Speaker A			Speaker B			Speaker C			Speaker D	
Meeting	b	c	d	b	c	d	b	c	d	d	
Alignment	42	60	-	130	104	87	89	52	-	-	564
Non-alignment	29	31	-	101	101	13	42	31	-	-	348
Subtotal	71	91	-	231	205	100	131	83	-	-	
Total	162			536			214			-	912

Interactional analysis of the 912 target turns resulted in 564 instances of alignment and 348 instances of non-alignment. In the two meetings (b and c), where participants A, B and C were all considered, B produced more than double the number of target turns (436) than were produced by C (214) or A (162).

As discussed earlier, the overall structure of the meetings is that Participant A and C report on the progress of projects to Participant B who has a more senior position compared to the other two participants. This might have influenced the organisation of turn sequences and may have led more often to Participant B being the second speaker in an adjacent turn sequence. Table 2 shows how often each speaker produced the domain IP, i.e. how often a speaker was "prior" speaker and how often the target IP, i.e. how often a speaker was the target speaker.

Not every speaker is followed by the other speakers equally often. It is notable that Participant C is rarely followed by Participant A (42 times), compared to 258 times he is followed by Participant B. Similarly, Participant A is followed by Participant C only 26 times, but by Participant B 178 times. The distribution of Participant B's contributions is more balanced, following A 120 times and B 188 times. This may be attributable to the overall characteristics of the research meetings, such as the different role occupied by B.

Table 2: Prior speaker and target speaker cross-tabulation.

Prior Speaker	Target Speaker			
	A	B	C	all
A	-	178	26	204
B	120	-	188	308
C	42	258	-	300
all	162	436	214	812

4.3.2 Inter-annotator agreement

In order to investigate if the sequential details described above can also be recognised by further analysts and used in order to classify alignments and non-alignments consistently, a second annotator was trained in the annotation scheme. This annotator was a masters student who has already gained some experience in doing CA. She had not been involved in the project at all before the point of classification. Some training was needed in order to make her familiar with the definitions of alignment and non-alignment

4.3.2.1 Training and categorisation task

The training data were taken from the first half of the meeting EN2009b.The training is based on eight extracts of orthographic transcripts containing 18 instances of target turns (see Appendix B). For each instance, the interactional category was suggested by the author, and discussed with the second annotator together with the observations that led to his decision. The training took almost one hour. After the training, the second annotator was given orthographic transcripts with the target utterances highlighted. The transcripts of the test set comprise are taken from the beginning of meeting EN2009b. Those, which were used in the training set, were excluded, resulting in 205 instances.

According to the instructions, the task was to classify each target utterance into one of the two interactional categories. Additionally, the annotator was asked to indicate on a three level scale from 0 to 2 how confident the classification decision was, where "2" means "confident", "1" means "fairly sure" and "0" means "just guessing" (the annotator used the expression "intuitive guess" for the value "0", when reporting the classification results). Information about confidence level had not been recorded for the first annotator.

For the categorisation task, the annotator was given one week to complete the task. The annotator could also choose the order of the transcripts and could come back to transcripts for comparison. She reported that the task took her 3.5 hours.

4.3.2.2 Inter-rater reliability

The classification of instances into the two categories alignment and non-alignment is summarised in Table 3.

Out of the 205 instances, the first annotator (R1) found 125 alignments and 80 non-alignments. The second annotator (R2) found 109 alignments and 96 non-alignments. In 99 cases both agree about alignments and in 70 cases both agree about non-alignments. In 26 cases, where the rater R1 chose the category alignment, R2 decided for non-alignment. In 10 cases, where the rater R1 chose the category non-alignment, R2 decided for alignment.

Table 3: The absolute and relative numbers of the categorisation of the two annotators.

First Annotator (R1)	Second Annotator (R2)		
	alignment	non-alignment	both
alignment	99 (0.48)	26 (0.13)	125 (0.61)
non-alignment	10 (0.05)	70 (0.34)	80 (0.39)
both	109 (0.53)	96 (0.47)	205

One way of comparing the ratings from two annotators is to report the percentage of agreement. This is the ratio between the number of ratings for which both raters agree and the total number of ratings. In this case it would be

$$Percentage_{agreement} = \frac{99 + 70}{205} = 0.82$$

However, this is of limited value as a measure of inter-rater reliability, since the raw percentage of agreement can be misleading: it is possible that two raters achieve high percentage agreement even if their ratings are random. By only looking at the observed percentage of agreement, one ignores the proportion of agreement that would be expected by chance, i.e. if the raters were scoring at random.

A better measure for inter-rater reliability is Cohen's Kappa (Cohen, 1960), (Wood, 2007). It adjusts for the "expected percentage of agreement". The formula for calculating Kappa is:

$$Kappa = \frac{(observed\ percentage_{agreement}) - (expected\ percentage_{agreement})}{1 - (expected\ percentage_{agreement})}$$

The "observed percentage of agreement" is, as calculated above: 82%.The percentages for those cases, for which the raters did not agree, are 5% and 13%.

The "expected percentage of agreement" is calculated in three steps. First, the marginals of the percentages have to be calculated by summing the percentages of alignments and non-alignments for each rater. Second, the appropriate marginals are cross-multiplied. Third, these two cross-products are added to get the expected percentage of agreement. This is the proportion of agreement between R1 and R2 that we would expect, if the two raters scored randomly.

Fed into the formula above,

$$Kappa = \frac{0.82 - 0.5}{1 - 0.5} = 0.64$$

On a scale from -1 (perfect and consistent disagreement) to +1 (perfect agreement), a value of 0.64 is in the range of "substantial agreement" (from 0.61 to 0.80). Lower values between 0.41 and 0.60 would be "moderate agreement" and values above 0.81 are considered as "almost perfect agreement" (Landis & Koch, 1977).

Kappa for different confidence levels

The second rater was asked to indicate a level of confidence for each decision. If the nine cases for which the rater had to make an "intuitive guess" are ignored, Kappa increased from 0.64 to 0.68. Additionally, if the other 54 cases, for which the rater was only "fairly sure", are also ignored, 142 "confident" decisions remain with a Kappa value of 0.78. The observed percentages of agreement are as shown in Table 4.

While the first annotator (R1) decided more often for alignments than for non-alignments, the second annotator (R2) has a balanced ratio of alignments and non-alignments (71:71). One possibility is that the second annotator was influenced by a perceived requirement to balance the decisions for the two categories. However, the instructions that she had been given did not mention that the categories might be equally balanced or that one category was expected to occur more often than the other.

Table 4: Absolute and relative numbers of the categorisation of the two annotators with the confident decisions of annotator R2 only.

First Annotator (R1)	Second Annotator (R2)		
	alignment	non-alignment	both
alignment	67 (0.47)	12 (0.09)	79 (0.56)
non-alignment	4 (0.03)	59 (0.42)	63 (0.44)
both	71 (0.50)	71 (0.50)	142 (1.00)

Discussion

There is an issue that Kappa faces, as summarised by Krippendorff (2004): "When the reliability of data is the issue, choosing κ is simply wrong for what it does. Its behaviour clearly invalidates widely held beliefs about κ, which are uncritically reproduced in the literature." (p. 421). The marginals seem to be causing the problem and Krippendorff suggests to follow Zwick's (1988) advice "to users of κ [...] to test for unequal margins before applying κ." (Krippendorff, 2004, p. 422).

However, looking at the marginal frequencies occurring in the current dataset, the small differences (125:109 and 80:96) indicate that this issue does not strongly apply here. Although there is considerable discussion on whether Cohen's Kappa or other inter-rater reliability

measures are reliable for normal distributions and for special distributions or not, the major trends in the current data are apparent: both annotators tend to agree on the two categories alignment and non-alignment, while the agreement on the non-alignment category is less than on the alignment category.

The inter annotator reliability score achieved for the categorisation into alignments and non-alignments can be compared with other researcher's results on similar categorisation tasks.

In work by Kurtić (2012), the raters had a similar binary decision task and had the choice between "competitive" and "non-competitive" overlap. Following Krippendorff, Kurtić reported an agreement of 0.62.The scored values found here between 0.64 and 0.78 (depending on the second speaker's confidence) are therefore comparable with those of Kurtić(2012, p. 121).

4.4 Discussion

In this chapter it has been shown that basic aligning and non-aligning actions can be distinguished through close analysis of sequential organisation and of verbal and non-verbal context. Comprehensive detail is provided that can be used to warrant the claim that a turn is aligning or non-aligning. Aligning actions are identified when the successive speaker is allowed by the second speaker's turn to continue on the prior speaker's agenda both, verbally and gesturally. Non-aligning actions are identified when this continuation on the prior agenda is actively inhibited by the second speaker's turn. The prior speaker successively orients towards the second speaker's turn, or the second speaker directly starts a new course of action with that turn. This answers research question RQ1 and supports previous interactional phonetics research, proposing such or similar fundamental actions (Szczepek Reed, 2012a; Wells, 2010; Stivers, 2008; Barth-Weingarten, 2011). It expands this previous work on aligning and affiliation with substantial work on the sequential and multimodal detail.

A list of criteria was drawn up which gathers cues that helps to explain to CA annotators the basic concept of the social actions "alignment" and "non-alignment" in order to train them to do a classification task. In that task, annotators were asked to identify whether the social action performed by the other speaker was aligning or non-aligning, i.e. whether the prior speaker's agenda was successively continued or inhibited. It refers to sequential, lexical, syntactic and semantic, but not prosodic cues (see Section 4.2).

Inter-rater reliability was measured with Cohen's Kappa and indicated substantial agreement between the two annotators. The findings suggest that alignments and non-alignments are distinguishable on the basis of the information provided in detailed orthographic transcripts alone, i.e. without reference to the additional phonetic and visual information available in the audio and video recordings.

It may be questioned whether this approach of testing inter annotator agreement involved circularity, as the second annotator was given some hints one could refer to order to make a classification decision. However, the training was merely used to make the annotator familiar with the definitions of the social actions "alignment" and "non-alignment". Similar training is necessary to introduce concepts such as "competitive" and "non-competitive" overlap (Kurtić, Brown, & Wells, 2009) to lay annotators. The agreement measure is not used to claim that specific sequential detail distinguishes the two actions – this is warranted by the close analysis of the interaction. The agreement measure is used to claim that it is possible to train other annotators in order to come to the same decisions. It shows that the details explained can be recognised by others and used in order to make a decision.

Although the classification of adjacent turns into alignments and non-alignments was successful, it is legitimate to ask if the concept of alignment is a dichotomy as it is presented here. One participant might choose to align with the other participant to a certain degree. Alignment could then be organised on a gradual scale from weak alignment to strong alignment. This might also be influenced by other actions that are performed in parallel or in the direct surrounding. For example agreements and disagreements (Ogden, 2006) may be related to the

organisation of interactional alignment. It could be argued that understanding checks (Kelly & Local, 1989) are a sub-class of non-alignments, as they require a specific reaction from the prior speaker. That is, the prior speaker "responds in a way appropriate to the [other speaker] having done an understanding check which treats the prior utterance as problematic and requires clarification of what was said." (p. 279). Another action that might influence the aligning character of an utterance is the response of requests (Walker, 2012). It could be argued that granting and non-granting requests influences the way prior speakers treat them as aligning or non-aligning. The actions quotation and mimicry (Couper-Kuhlen, 1996) may also be used by participants in order to align or disalign with the prior speaker. Alignment and non-alignment might also be influenced at places where participants overlap and either compete for the floor or not (Kurtić, Brown & Wells, 2009; Kurtić, 2012).

In the next chapters, the prosodic analysis and gestural analysis of the adjacent turns from the two speakers are described.

5 Acoustic analysis

The interactional analysis was purely based on the orthographic transcripts including gestures, but without reference to prosody. They contain the individual turns of speakers, expressed verbally (words) and non-verbally (gestures). Going one step further and listening to the aligning and non-aligning utterances and their sequential context gives the impression that most of the aligning turns match prosodically with the last intonation phrase of the prior speaker, while the non-aligning turns do not. Similar indications towards prosodic matching and non-matching can be found in the literature (Müller, 1996; Szczepek Reed, 2004, 2006; Walker, 2004; Couper-Kuhlen, 1996). Similar phenomena are also referred to as entrainment, synchrony, accommodation, and convergence by Wichmann (2011, unpublished): "In cooperative talk, there is a tendency for participants to remain "in tune" and "in time"." (Wichmann, 2011, Section 5).

As we have seen from the interactional analysis, speakers do not always "cooperate" in the same way. They can for example decide not to align with the prior speaker's agenda and make the prior speaker move away from the current course of action. Research to date has shown that there is evidence that this move is performed by the second speaker by prosodically non-matching the last IP of the prior speaker's turn.

According to the research questions on prosodic matching (RQ2a-c), the general aim of the acoustic study is to test whether the interactional categories depend on the prosodic match of the two adjacent turns. There is the possibility that an agenda aligning turn is prosodically similar to the preceding turn. In other words, an utterance which is prosodically similar to its predecessor seems to do agenda aligning work. Therefore we need to develop a way of comparing intonation contours according to acoustic parameters that are responsible for the identification of prosodic matches and non-matches (RQ2c). Such a method should also be objective (RQ2b).

5.1 Prosodic similarity metrics

Two metrics are devised in this chapter that describe alternative metrics to the ones suggested by Hermes (1998a, 1998b) and by Rilliard et al. (2011), used to determine the similarity of utterances in terms of prosody. They refer to prosodic features from two utterances produced by two different speakers. One basic requirement is that they take both into account, the movement of the F0 contour and the range in which this movement occurs. The first metric (Section 5.2) has been presented in an article by Gorisch, Wells and Brown (2012). It evaluates the similarity of contours of different durations using the "maximum similarity score" within time windows. Below, we explain this technique, evaluate it and expose any shortcomings. Thereafter we propose a second metric that employs dynamic programming to generate an "accumulated quality score" for the overall similarity of the two utterances (Section 5.3).

As discussed above, the spontaneous conversational speech that we analyse in this thesis is dominated by relatively sparse F0 readings. Hermes' and Rilliard et al.'s data represent recordings of clear speech from auditory experiments. Additionally to analysing "clean" speech, their technique interpolated the remaining gaps in the F0 contours, due to voiceless regions. Our premise, however, is to avoid interpolation for the missing data regions, as the gaps can be rather wide.

As both metrics require some processing of the acoustic signal, this will be explained in the following.

5.1.1 F0 extraction and normalisation

Couper-Kuhlen (1996) showed that for participants in talk-in-interaction it makes an interactional difference whether the second speaker matches the first speaker's contour on a relative or an absolute pitch register. Individual F0 normalisations for every speaker are therefore needed. F0 normalisation allows for two parameters: first, location of each speaker's F0 contour within the individual overall range; and second regarding the F0 span: information on how far away from his or her mean F0 the speaker's contour falls or rises. The normalisation of the F0 parameter is an essential step in determining how to objectively measure prosodic similarity (RQ2b).

5.1.1.1 F0 extraction

The F0 of each of the three speakers was computed over the entire duration of every meeting. We used the YIN pitch detection algorithm (De Cheveigné & Kawahara, 2002), which has the option to retain only those stretches of the F0 contour which coincide with high periodicity in the signal (by setting a threshold of aperiodicity, which here was set to 0.2). The choice of pitch estimation is not crucial here as far as it provides a means of specifying a threshold of voicing or another technique that allows to determine those values, which count as belonging to the fundamental frequency and those values, which are not part of that feature. We would expect similar results using other pitch determination algorithms, such as the one provided in Praat (Boersma & Weenink, 2012). An evaluation of different pitch extraction algorithms has shown little difference in the quality of resulting F0 contours (Kurtić, 2012, p. 172 f.).

5.1.1.2 F0 normalisation

The distributions of all F0 values obtained with the YIN algorithm are displayed for the individual speakers in Table 5 (speaker A in the first row; speaker B in the second; speaker C in the third; speaker D in the fourth row). The F0 values are scaled logarithmically in order to roughly approximate the subjective measure of pitch (Plack, 2005). Some of the distributions, especially the distribution from Speaker A show multiple peaks: one peak corresponds to the values from the target speaker, and the others correspond to F0 values from the other speakers.

This phenomenon is due to crosstalk from the other two or three speakers that is picked up by the target speaker's microphone, which, although it has a low sound level, is able to influence the F0 distribution. The amount of influence of the signal from the other speakers on the distributions may depend on the headset microphones and their orientation. This problem was addressed by retaining the F0-contours only for those regions that coincided with speech from the target speaker, as identified from the word-level transcription. By doing so, F0 is estimated only when the voice of the target speaker is likely to exceed the level of crosstalk. Since YIN identifies the most dominant pitch period in the acoustic signal, reliable F0 tracks can then be obtained. The resulting distributions are displayed for the individual speakers in the mid column of Table 5 (speaker A in the first row; speaker B in the second; speaker C in the third; speaker D in the fourth). The data in the first two columns arise from meeting D, at which all four speakers were present. Similar results were obtained for the other two meetings B and C. Summary statistics of the F0 values found in meetings B, C and D are shown in Table 6.

The F0 distributions were skewed to higher F0 values, especially for the female speaker B (see second row in Table 5). The reasons for this effect are speculative; the way in which speakers make use of their F0-range may depend on the environment in which the conversation takes place, on the task in which the speaker is involved, and on other factors. Given the skew towards high F0 values, we take the median as the reference point of the speaker's mid range, as most of the produced values can be observed around the median (cf. Walker, 2004).

Along with the median, the variance of the data has to be taken into account. The amount of excursion from the median of one speaker may be different from the amount of excursion from the median of another speaker. In order to take this into account, we normalise the speaker's

distributions according to their standard deviation (cf. Heldner, Edlund and Hirschberg, 2010). Specifically, the F0 values were logarithmically scaled $\hat{f} = \log_2(F0)$ and then the normalised contour was obtained by $f = (\hat{f} - m)/s$ where m and s represent the median and standard deviation of $\hat{f}$ respectively. These F0 contours, normalised for speaker characteristics, provide the basis for comparison across speakers with the aim of identifying prosodic matches.

Table 5: Distribution of F0-values according to speakers A to D (in rows top - down) and according to different methods: all F0-values (left column), F0-values based on the word segmentation (middle column). The left and the middle column are based on F0-values from meeting D; the right column takes values from all meetings B, C and D. Multiple peaks can be seen in the left column (due to cross-talk). The separation from other speaker's talk is achieved by focussing on the word segmentation (see middle column). The distributions of the right column are used to specify the speakers' mid F0- range (median) and the variation in F0 (standard deviation).

Raw F0-values from meeting D	F0-values based on word segmentation from meeting D	F0-values based on word segmentation from all three meetings B, C and D

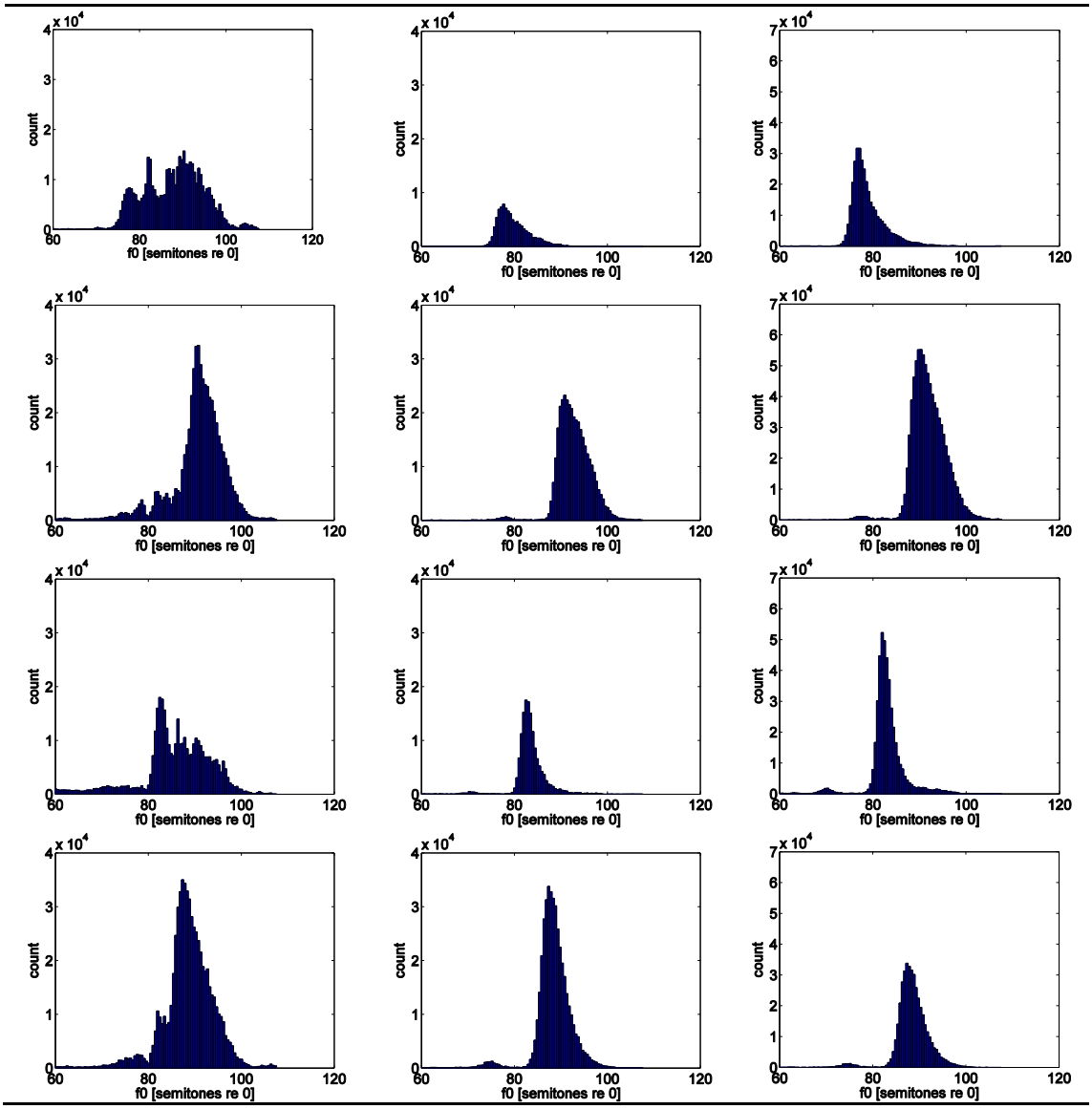


Table 6: Summary of statistical measures of the distribution of F0 values in Hz and semitones (re 0 Hz) according to speakers and meetings.

Meeting	Speaker	Mean		Median		Standard deviation	
		[Hz]	[st re 0]	[Hz]	[st re 0]	[Hz]	[st re 0]
EN2009b	A	92	78.09	85	76.99	25	3.95
	B	200	91.15	189	90.66	45	3.91
	C	121	82.80	115	82.29	28	4.68
EN2009c	A	98	78.99	91	78.04	26	3.61
	B	201	91.43	191	91.00	44	3.74
	C	124	82.90	118	82.64	33	4.35
EN2009d	A	104	79.94	96	79.07	31	3.89
	B	217	92.77	209	92.53	46	3.81
	C	127	83.38	121	83.11	32	4.05
	D	169	88.39	164	88.33	36	4.05
Across all meetings	A	99	79.01	91	78.01	29	3.85
	B	207	91.90	198	91.56	46	3.88
	C	124	83.03	119	82.71	34	4.38
	D	169	88.39	164	88.33	36	4.05

5.1.2 Intensity extraction and normalisation

Because the microphone channels may have been recorded with different levels, and because of the uncertainty about the distance of the microphone from the mouth and individual differences in the intensity of speaking, normalisation was also carried out for the intensity contours. The psychoacoustic equivalent of intensity is loudness. But it is not clear how far loudness is related to other perceptual quantities, as it is influenced by F0, voice quality, sound production, etc. Therefore at this stage we use the original intensity signal of the sound.

5.1.2.1 Intensity extraction

The intensity contour is computed from the instantaneous power of the signal, smoothed according to the fundamental period. This is also computed by the YIN algorithm. We then transform the intensity values into decibels. Because there is no information in the AMI-meeting corpus description about a standard tone used for calibrating the sound level, we chose an arbitrary value as reference. The recorded value of 0.000002 was assigned to 40dB.

In contrast to other work commonly referring to intensity contours extracted using Praat (Boersma & Weenink, 2012) large smoothing windows are not employed here. Such smoothing introduces relatively high intensity values at times where none or very little intensity is present in the signal, due to the proximity of surrounding high intensity regions.

5.1.2.2 Intensity normalisation

In the same way as for the F0 normalisation we take the distribution of all values of the intensity contour of each speaker over the whole meeting. Example distributions are displayed for meeting D according to the four different speakers: speaker A (Figure 1), speaker B (Figure 2), speaker C (Figure 3) and speaker D (Figure 4). Analogous to meeting D, the distributions of intensity values were obtained for the other two meetings B and C. In order to obtain the normalised intensity contour we subtract the median and divide the result by the standard deviation.

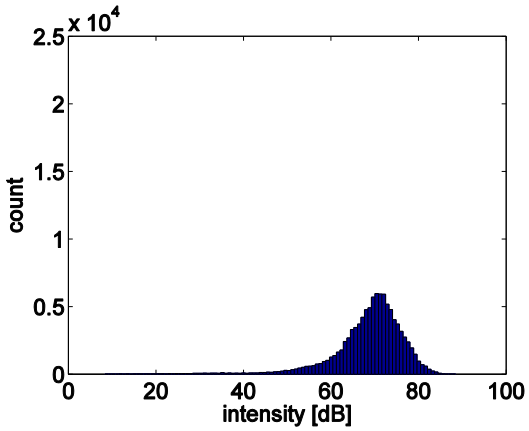


Figure 1: Speaker A; distribution of intensity-values.

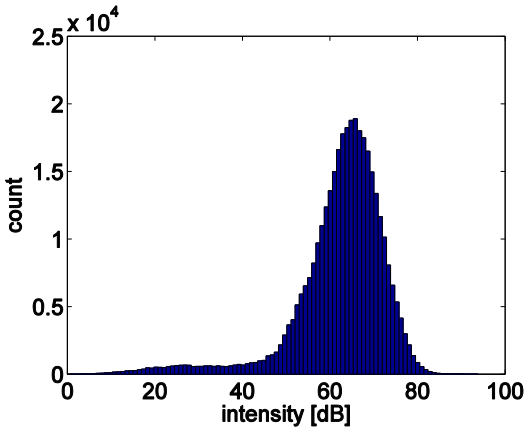


Figure 2: Speaker B; distribution of intensity-values.

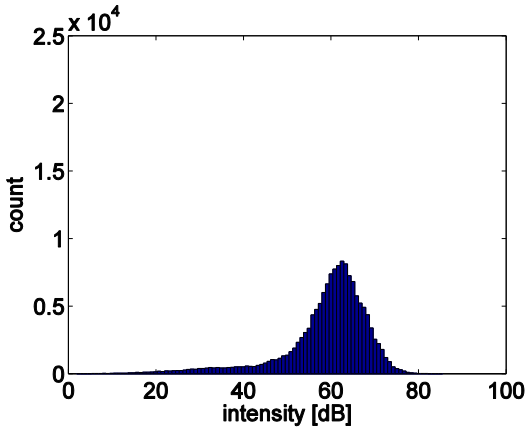


Figure 3: Speaker C; distribution of intensity-values.

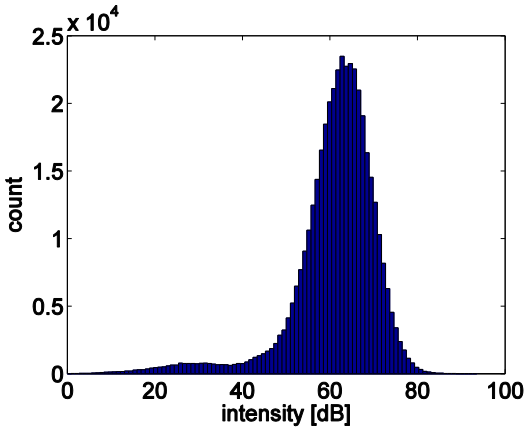


Figure 4: Speaker D; distribution of intensity-values.

In contrast to the normalisation of the F0 contours, where the distribution of F0-values were combined across all meetings in order to obtain the median and standard deviation, the normalisation of the intensity contours was performed separately for each meeting. We did this because of the variation of the experimental setup. Participants may sit at different positions with different headset microphones and the sound level may change with re-adjustments of the microphone placement and the recording level.

5.2 Maximum similarity score

The research question addressed in this study is how prosodic similarity can be measured objectively (RQ2b). One metric that measures the F0 similarity was presented by Gorisch, Wells and Brown (2012) and is reviewed here. It is based on a similar approach used by Cooke (1993) to compare amplitude modulation contours in a computational auditory scene analysis model. Given two instantaneous F0 values x and y, their similarity $sim(x, y)$ is computed as:

$$sim(x, y) = e^{\frac{-(x-y)^2}{2\sigma^2}} \quad (1)$$

Here, a Gaussian function is applied to the difference between the F0 values. When the difference between the two F0 values is small, i.e., when it lies on the broad peak of the Gaussian function, the similarity is close to 1. A large difference between the F0 values has a similarity of zero. In this way, the difference of the F0 values is converted into a concept of similarity. The parameter σ represents the width of the Gaussian function. Over this Gaussian function, the F0 differences receive a value between 0 and 1. Large differences are assigned the values close to 0 and small differences are assigned values close to 1. When σ is larger, more

value pairs with large difference receive a higher value. Further, Gorisch et al., set σ to a range of values (0.1, 0.2, 0.3 and 0.4). This value is not critical; qualitatively similar results were obtained across this range of σ values (see result section below).

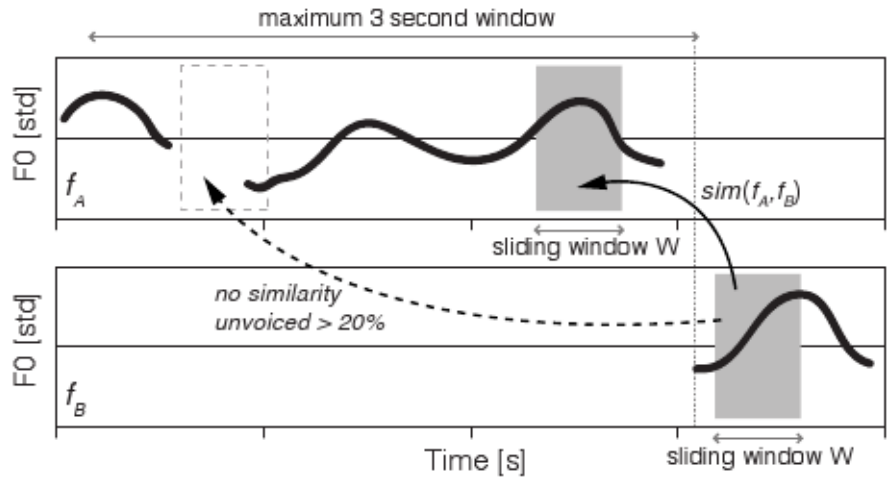


Figure 5: Comparison of two F0 contours f_A and f_B using a sliding window. Within each window (in grey), the F0-values are compared one against the other for their similarity ($sim(x, y)$). If the amount of unvoiced segments in a window exceeds 20%, the similarity for that window is not computed. The step-size from window to window is 1. From Gorisch, Wells and Brown (2012: 68) Language and Speech, 55/1, March/2012 by SAGE Publications Ltd., All rights reserved. © 2012 SAGE Publications Ltd. Reproduced by permission of SAGE.

The equation given above describes how they quantify the similarity of two instantaneous F0 values. However, their aim is to determine the similarity of two F0 contours, which will vary in duration. Accordingly, they adopt the scheme shown schematically in Figure 5, in which sections of the F0 contour f_B of speaker B's speech are compared with sections of the F0 contour f_A of speaker A's speech. For the present study, f_A constitutes the prior speaker's talk, which they previously called *prior turn* or *domain turn*; f_B represents the *current turn* or *target turn*.

The duration of f_A was limited to 3 seconds, which is consistent with Pöppel's (2009) suggestion that there is a time window of two to three seconds of "subjective presence".

5.2.1 Maximum similarity measure

Within the 3-second scope of the preceding speech, the F0 of the last intonational phrase of the prior speaker (domain IP) is compared with the F0 of the intonational phrase of the current speaker (target IP). Because of differences in duration of the prior speaker's IP and the target IP, sliding windows are used to perform the comparison. In practice, as it was argued above, there are voiceless parts or weak evidence of F0 due to voice characteristics such as creaky or breathy voice. As a result, it is necessary to find a compromise between window size and the percentage of regions where one or the other of the F0s is not available. To account for a reasonably long stretch of talk without introducing too many gaps in the F0 contours, Gorisch et al. used a sliding window with size W of 120 ms and accept no more than 20% of unvoiced time frames (i.e., if the proportion of unvoiced time frames exceeds 20%, then a similarity score is not computed). The similarity of f_A and f_B is computed as an average over the sliding window W, excluding the voiceless parts. More formally, they compute the F0 similarity as

$$sim_{AB} = \frac{1}{|V|} \sum_{t \in V} sim(f_A(t), f_B(t)) \quad (2)$$

Where V is the subset of time indices within W for which both $f_A(t)$ and $f_B(t)$ are available and $|V|$ is the cardinality of V. Due to the windowing over both F0 contours, sim_{AB} can be

represented as a two-dimensional similarity matrix (see Figure 7 below) in which the sliding window position within f_A is shown on the abscissa, the window position within f_B is shown on the ordinate, and the similarity value is represented by means of a grey scale.

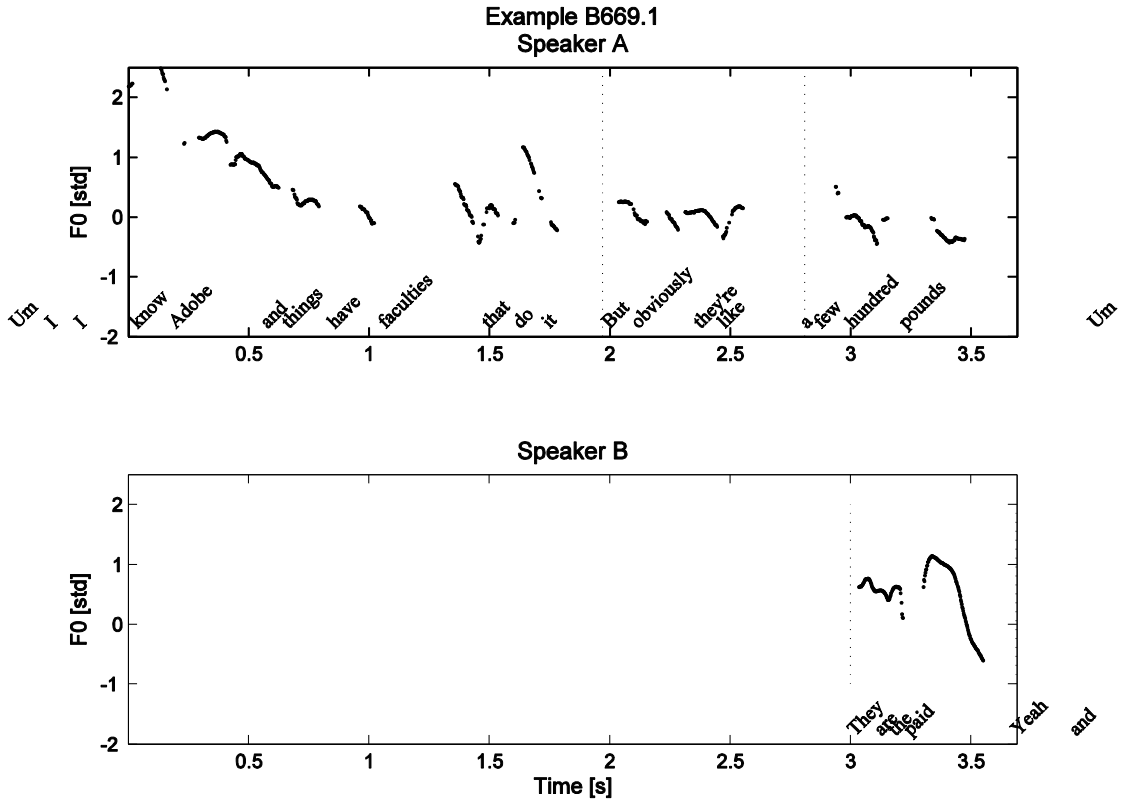


Figure 6: F0 contours of a turn sequence from Extract 11. Speaker A (top panel) holds the floor and is overlapped by speaker B (bottom panel) starting at the time step of three seconds. The two turns (IPs) of interest are speaker A's "but obviously they're like" (delimited by the vertical dotted lines) and speaker B's "they are the paid". Speaker A's F0 is slightly above the mid range and varies with half a standard deviation. Speaker B's F0 is a little more above the speaker's mid. On the word "paid", the F0 first rises to one standard deviation above the speaker's mean and then falls one and a half standard deviations.

In order to measure the prosodic similarity between the domain IP (prior speaker) and the target IP (current speaker), Gorisch et al. take the maximum similarity score within the cross-section of the two turns in the similarity matrix.

$$max_{sim} = \max\left(sim_{AB}\right) \tag{3}$$

where *A* and *B* are delimited by the start- and end-times of the according IPs.

The procedure is now illustrated according to an example from the real recordings. We use an interactional sequence from Extract 11, where speaker B overlaps speaker A's utterance. Figure 6 displays the F0 contours of the two speakers.

The similarity sim_{AB} of the F0 values, windowed over both utterances, is displayed in Figure 7. The higher the similarity between the windowed contours, the darker is the corresponding point in the matrix. In the current example, the similarity matrix comprises the comparison of speaker B's utterance "they are the paid" (0.7 seconds long) with the entire preceding and overlapped utterance of speaker A "I know Adobe and things have faculties that do it but obviously they're like a few hundred pounds" (3.7 seconds). The area of interest is here the last IP of speaker A ("but obviously they're like"), delimited by vertical dotted lines, before speaker B's IP ("they are the paid"). In this example the maximum similarity max_{sim} within this area is 0.7.

Explained in more linguistic terms the above means that the IPs are compared according to their F0 values. Each F0 value from one contour is compared with each F0 value of the other contour. First, the difference between the individual F0 values is computed. Second, this difference score is assigned a similarity score according to the function expressed in Equation (1). This simply turns a F0-difference (between infinity and 0) into a F0-similarity (between 0 and 1). A similarity score close to 1 means that the two F0-values are very close. A similarity score close to 0 means that the F0-values are very far apart. What can be seen in Figure 7 is a matrix that includes all pairwise F0 comparisons. The grey patches indicate how similar the corresponding regions of the F0 contours from the two speakers are: darker means more similar while lighter means more dissimilar. Those regions for which not enough readings of the F0 contour were available for a reliable comparison are left completely white.

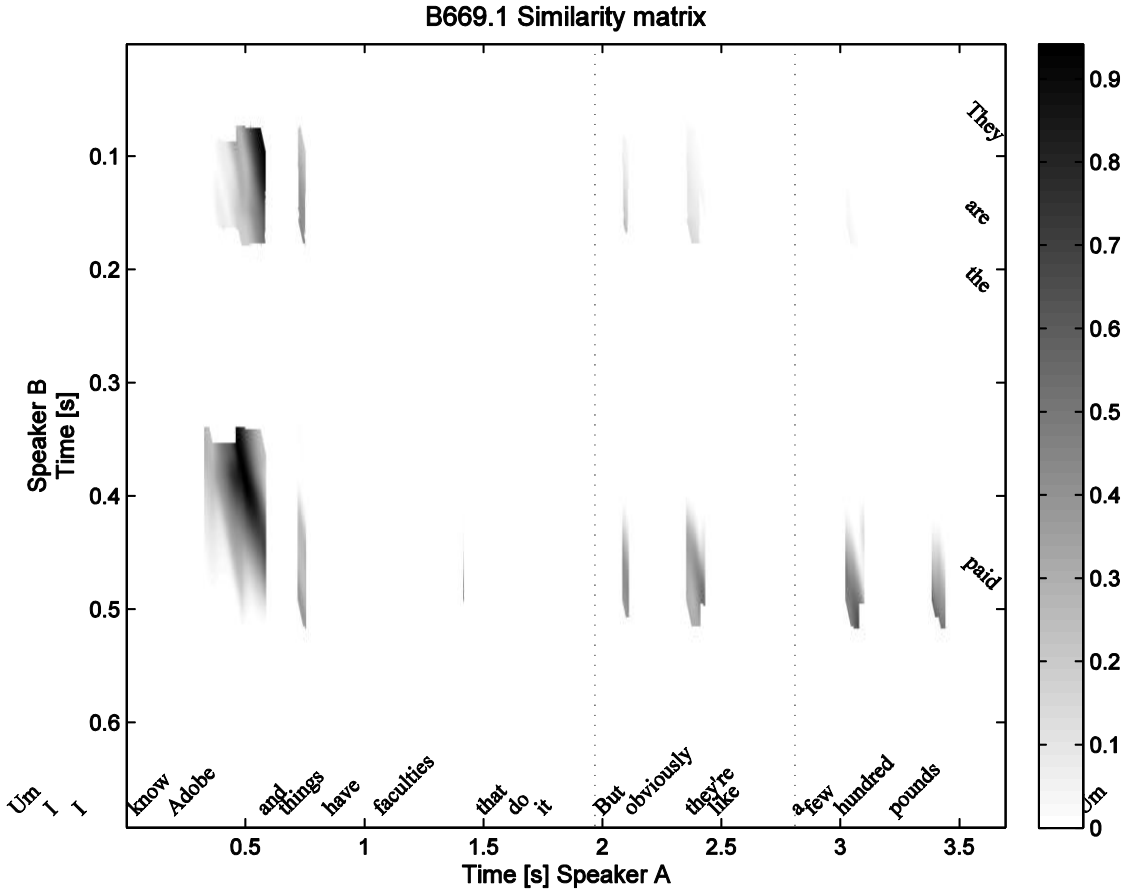


Figure 7: Similarity matrix of the F0 contours from Figure 6 of speaker A's utterance (on the abscissa from left to right) and speaker B's utterance (on the ordinate from top to bottom). The similarity is indicated on a grey scale from white (low similarity) to black (high similarity). The highest similarity (above 0.9) between the two utterances is at around 0.6s of speaker A's utterance, outside of the two IPs "but obvious they're like" and "they are the paid". The maximum similarity within the two IPs is 0.7.

Thus where an utterance is compared to an exact copy of itself, one would see a black diagonal line from the top left corner to the bottom right corner of the similarity matrix, as pairs of identical F0 values will always appear on this diagonal. This by definition results in a difference of 0, which is a similarity of 1. Depending on the shape of the contours, the rest of the matrix would be filled according to the similarity scores of the pairwise comparisons. They can range from 0 to 1 with the corresponding shades of grey, except for the voiceless regions, where the colour would be white.

5.2.2 Intensity weighting

Arguably, not all regions of the two F0 contours should be given the same weight in a matching comparison. F0 values that occur in relatively intense regions of speech should be given higher weight (Harris, 1947; Hermes, 1998b; Rilliard et al. 2011).

The intensity contours of the current example (from Extract 11) are displayed in Figure 8. It can be observed, that the intensity varies rapidly over the utterances. For example the "`the`" in speaker B's utterance "`they are the paid`" is produced with relatively low intensity (close to zero) compared to the surrounding words.

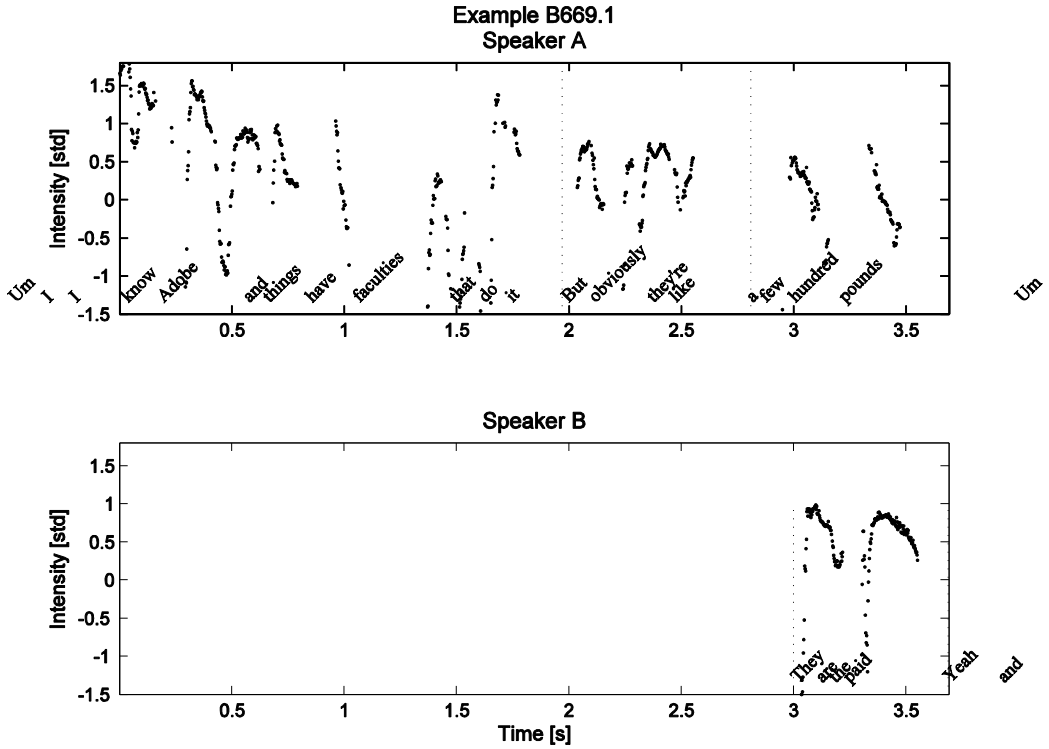


Figure 8: Intensity contours from the example in Extract 11. Intensity is normalised for median and standard deviation. The intensity in speaker A's utterance (top panel) and speaker B's utterance (bottom panel) fluctuates rapidly due to closure and bursts of plosives and different levels of voiced sounds.

In an alternative display, the intensity information is used to weight the F0 contours as is shown in Figure 9. The amount of extension of the F0 contour in the vertical direction is directly proportional to the intensity with which the individual F0-values have been produced. This display, which includes two prosodic parameters in one visual representation, was developed in relation to the hypothesis that intensity weighting is a useful method in the analysis of F0 contours (Gorisch, 2010). It aimed at helping the prosodic analyst to have an idea of the parameters at one glance. This was presented at the Colloquium of the British Association of Academic Phoneticians (BAAP2010) and finds an adaptation in work by Walker (2012), who uses a dotted line for F0 that is modulated on a grey scale according to intensity.

Here, the F0 contour extends along the vertical axis according to intensity. Thereby one could get the impression that the line was made fatter and that one would lose some accuracy of F0. If the line was simply made fatter, this would be counterintuitive: the louder something is, the fatter would be the line, which means the less accurate F0 would be visually, although in fact the F0 would be perceptually more salient when the intensity is high. However, the F0 contour is not simply made fatter, which would affect both dimensions: vertical and horizontal, but extends only in one dimension, the vertical one. This way, no information of F0 is lost, because the actual F0 always corresponds to the mid-point of the patch. This might not be the optimal visual representation. For example in regions of fast F0-movements of equal intensity, the line

is a bit thinner than in regions of static F0. It means that there is space for improvement, but the display is at least one that integrates both F0 and intensity. A similar display is used by Neiberg, Salvi, & Gustavson (2013), who use in fact the line width for modulation, extending in both dimensions vertical and horizontal. Possibly it needs some time for researchers to adapt to such representations in order to work with them effectively.

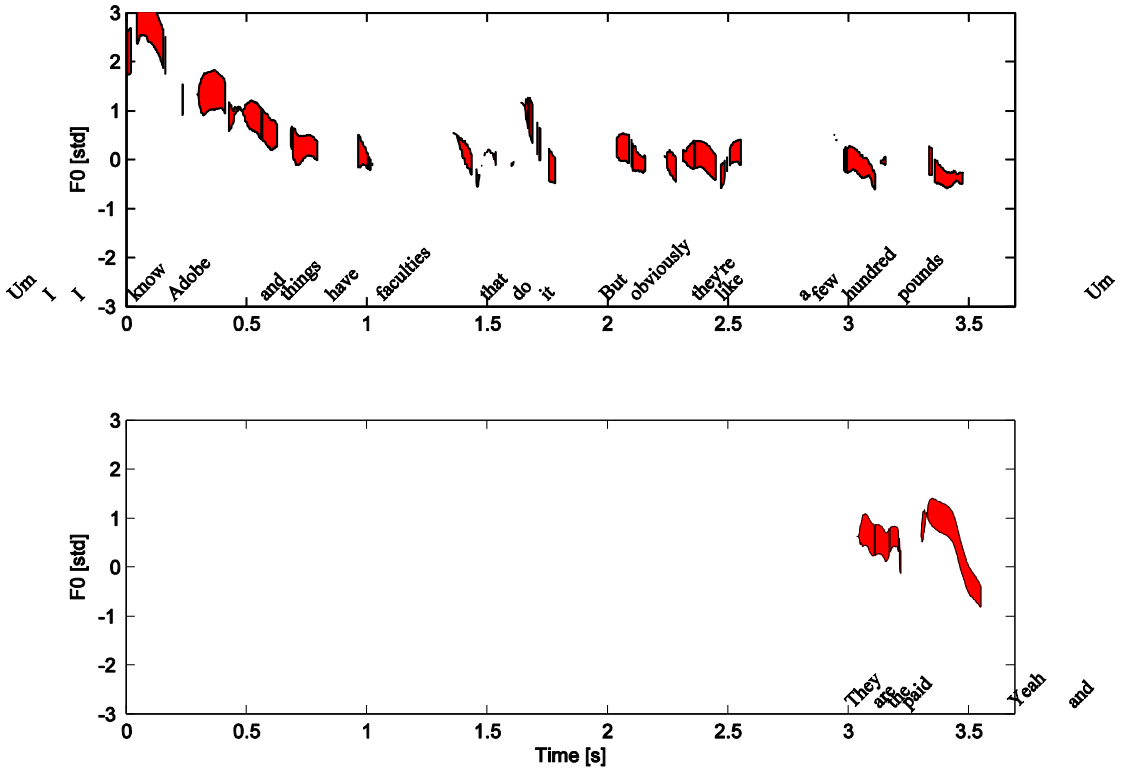


Figure 9: Intensity weighted F0 contours from the example in Extract 11. F0 contours (cf. Figure 6) from speaker A's utterance (top panel) and speaker B's utterance (bottom panel) extend vertically (red) with the intensity by which they are produced (cf. Figure 8). F0-values of the F0-contour that are produced with low intensity extend less in the vertical direction, while F0-values that are produced with high intensity extend more. For example the rising start of the F0-contour on the word "paid" by speaker B is produced with much less intensity than the flat middle and the falling end of the F0-contour. This information, which is not contained in the raw F0-contour illustrated in Figure 6 is added here.

Accordingly, Gorisch et al. (2012) also employ an intensity-weighted version of the similarity metric. When the average intensity of the two talkers is low, they expect that any difference between the two pitch contours should contribute less to their similarity. The average intensity is given by

$$\alpha(t) = \begin{cases} A(t), & if A(t) > 0 \\ 0, & otherwise \end{cases} \quad (4)$$

Where

$$A(t) = \frac{I_A(t) + I_B(t)}{2} + c \quad (5)$$

The constant c is added to ensure that the majority of $\alpha(t)$ values are positive (the intensity values are in decibel units, so that they roughly correspond to perceived loudness; they have been normalized according to the median and the standard deviation of each speaker's intensity). Occasionally, values smaller than c occur, in this case they are clipped to zero. The intensity-weighted F0 similarity metric (normalized by alpha) is then given by

$$iwsim_{AB} = \frac{\sum_{t \in V} \alpha(t) sim(f_A(t), f_B(t))}{\sum_{t \in V} \alpha(t)} \quad (6)$$

The resulting similarity metric tells us how close the similarity of the F0 contours is in terms of F0 movement and F0 height. However, if the mean intensity across the two speakers is low at a particular time instant t, the F0 similarity at that time contributes less to the overall similarity.

Above we describe a technique that is intended to estimate the prosodic similarity between two utterances. It was introduced by Gorisch, Wells and Brown (2012) and was applied to data from real recordings of interacting participants. Results showed that aligning short turns are on average more similar in terms of prosody than non-aligning turns. It is a valuable contribution to the field of interactional phonetics as it demonstrated an instrumental approach of testing the prosodic matching hypothesis. However, the underlying data collection (177 instances) was relatively small and the relation between alignments (149 instances) and non-alignments (28 instances) was very disproportionate. A return to this metric with an increased and more balanced data-collection was intended in that paper in order to either confirm or disconfirm the findings. The results on a larger data set are presented in Section 5.4.

5.2.3 Evaluation with artificial contours

Although the maximum similarity search algorithm revealed the differences in prosodic matching between aligning turns and non-aligning turns, Gorisch et al. (2012) didn't present a detailed analysis of how well the algorithm estimates prosodic similarity. It is sensible to evaluate its capacity by using artificial contours for which we can make predictions of similarity. This way it is possible to check that the model actually works as expected.

To our knowledge, no standardised set of contours exists for which prosodic matching algorithms can be applied and evaluated. In the current study we propose such a set of artificial contours. We attempt to keep the shape of the contours as close to the shapes that are observable in real F0 contours. Thereby we maintain the possibility to formulate expectations on their similarity and to make qualitative predictions about the results of the algorithms.

In order to test the maximum similarity search algorithm and the DTW algorithm (which is described in Section 5.3), different types of artificial contours are created. They are parts of simple sine waves with specific characteristics. Modulation of the duration (time), the frequency (slope), height (range) and the introduction of gaps (missing data, voicelessness) makes it possible to construct pairs of first and second contours, which are related to the phenomena which can be observed in the real data. However, the following experiments are performed using controlled data. We make several demands on the algorithms as follows:

Contours which are different (e.g. in duration, frequency, height), should get a lower overall similarity score than contours which are largely similar.

Contours which have the same overall shape (e.g. rise-fall), but which are stretched in time should get the same similarity score. For example, it should not make a difference if a short "`yeah`" matches a short "`da`" or a relatively longer utterance "`da da da`", as long as the overall shape (e.g. rise-fall) is the same.

Gaps in at least one of the two contours should not substantially decrease the overall similarity score if the underlying contours are the same. Thereby neither the size of the gaps nor the location of the gaps should influence the similarity score.

5.2.3.1 Artificial contours

Each of the parameters can be modulated dynamically. Starting from the same contours (first pair in a type) and modifying the parameters step by step, the different contour types should be categorised as follows:

a) ConstTimeDiffFreq: constant time; modulated frequency (Figure 10)

b) DiffTimeConstFreq: distorted time; constant frequency (Figure 11)

c) DiffTimeDiffFreq: both: distorted time and modulated frequency (Figure 12)

d) ConstFreqConstTimeDiffRange: constant frequency and time; modulated height (Figure 13)

e) ConstFreqOneGap: constant time and frequency; gaps at different time stretches in second contour(s) (Figure 14)

f) ConstFreqTwoGap: constant time and frequency; one gap in first contour; gaps at different time stretches in second contour(s); one time stretch coincides with the time stretch of the first contour (Figure 15)

g) ConstFreqOneGrowingGap: constant time and frequency; gaps of different size in second contour(s) (Figure 16)

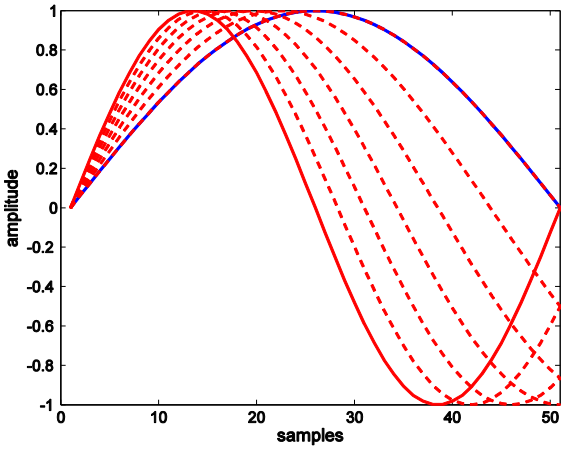


Figure 10: ConstTimeDiffFreq: One first contour (blue) and seven second contours (red). The time is kept constant while the slope (frequency) in the second contours increases and reaches the double of the first contour. One of the red contours is identical with the blue contour.

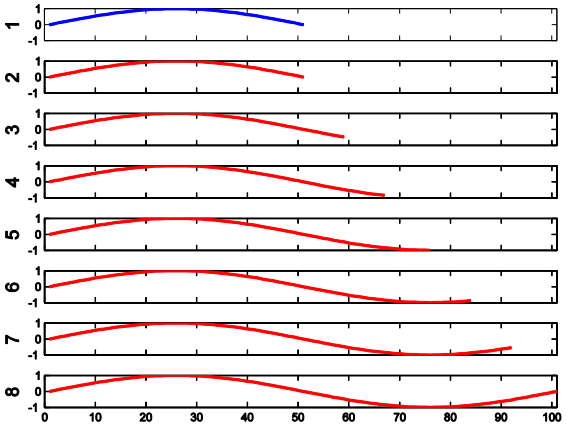


Figure 11: DiffTimeConstFreq: The frequency is kept constant while the time increases to the double.

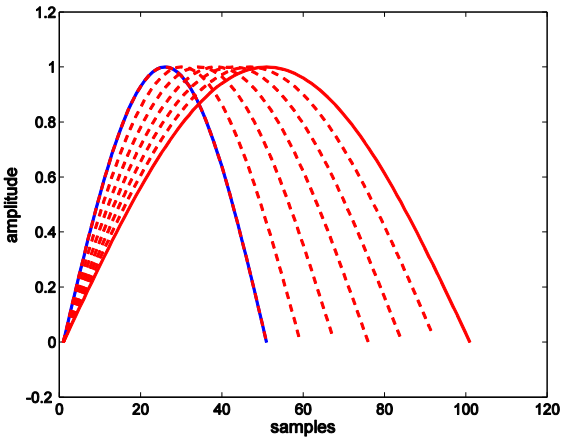


Figure 12: DiffTimeDiffFreq: Time increases to the double value and frequency decreases to the half value.

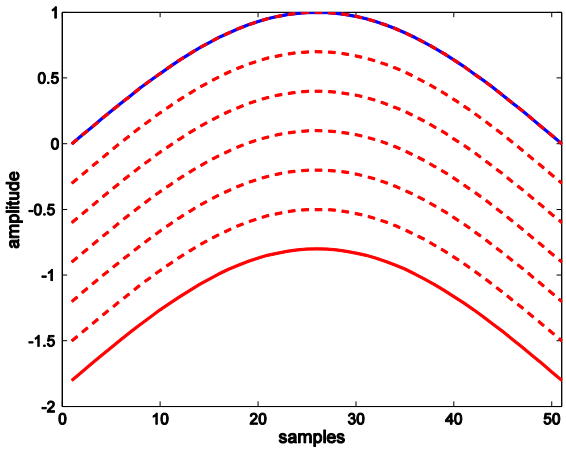


Figure 13: ConstFreqConstTimeDiffRange: Time and frequency are kept constant while the height of the second contours decreases from 0 to -1.8 in steps of 0.3 relative to the blue contour.

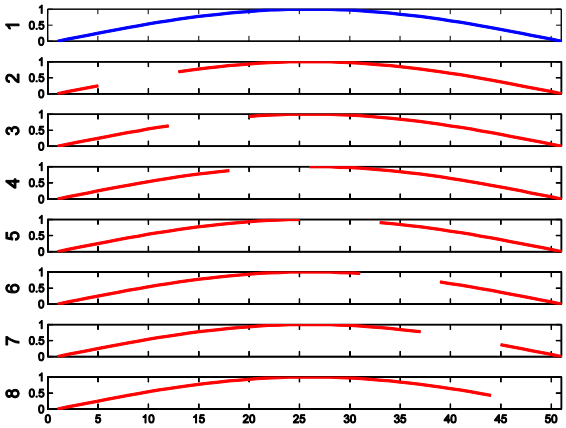


Figure 14: ConstFreqOneGap: Time and frequency are kept constant while in the second contours a gap is introduced at different phases.

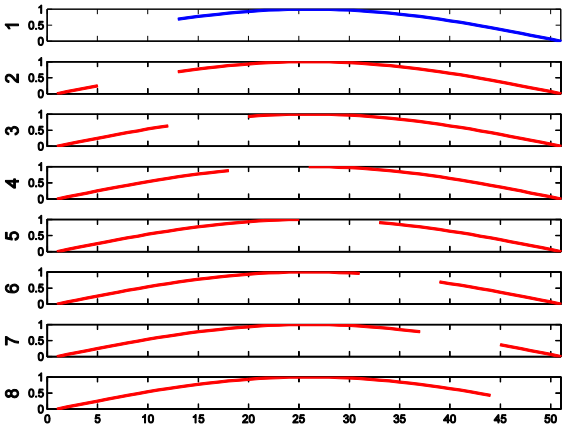


Figure 15: ConstFreqTwoGap: Time and frequency are kept constant while one gap is introduced to the red contour at different phases and one gap is in the first contour. At a certain phase the two gaps coincide.

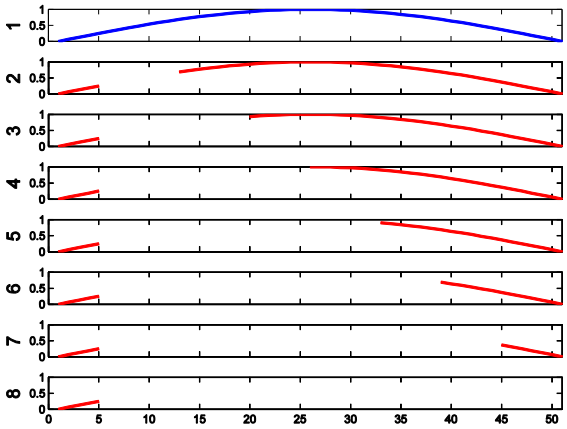


Figure 16: ConstFreqOneGrowingGap: Time and frequency are kept constant while one gap of different size is introduced in the second contours (red).

5.2.3.2 Contrasting artificial contours

In the set of contours described above some of the red contours are already quite different from the blue line. These differences result from modifying the parameters of the underlying sine-waves frequency, time and height and from the introduction of varying gaps. But the initial contours are always the same – characterized by a rise-fall shape. A second set of contours are devised below where even the initial contours are already different – the rise-fall contour shape is contrasted with the fall-rise contour shape, building a set of basically "inverted" contours.

Starting from contrasting "inverted" contours (first pair in a type) and modifying the parameters listed as above step by step, contours result, which are displayed in Figure 17 to Figure 23:

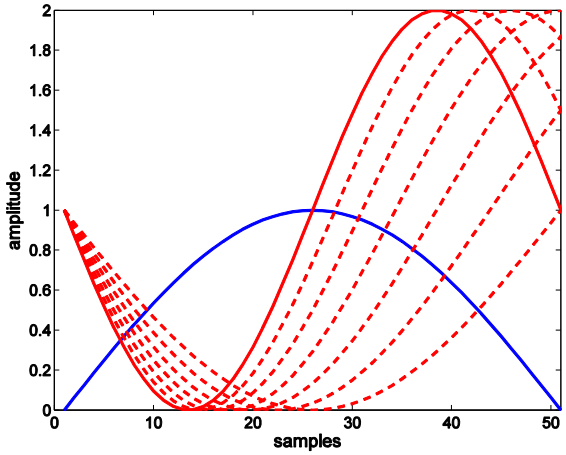


Figure 17: ConstTimeDiffFreq: A rise-fall contour is contrasted with a fall-rise contour. Successively the frequency of the basic sine wave is increased and has finally (solid red line) a frequency which is double of the initial frequency.

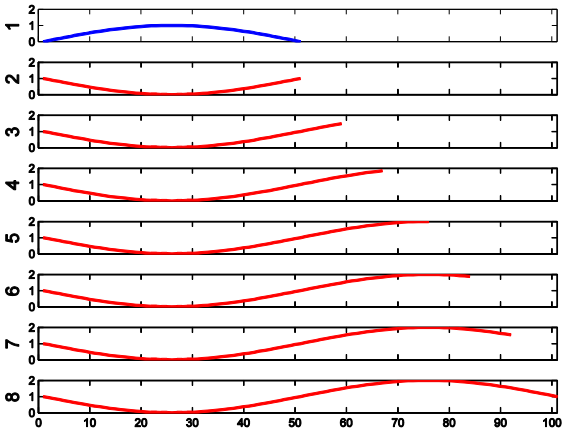


Figure 18: DiffTimeConstFreq: The frequency is kept constant while the time interval increases, which has finally doubled its duration. It has at its extreme a fall-rise-fall contour.

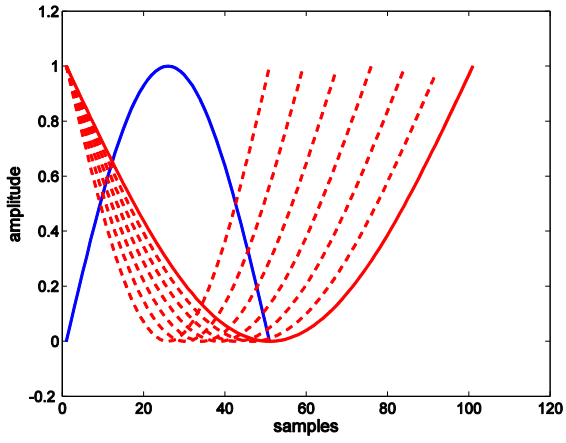


Figure 19: DiffTimeDiffFreq: A rise-fall contour is contrasted with contours which have a fall-rise shape. This shape is preserved while the time is stretched.

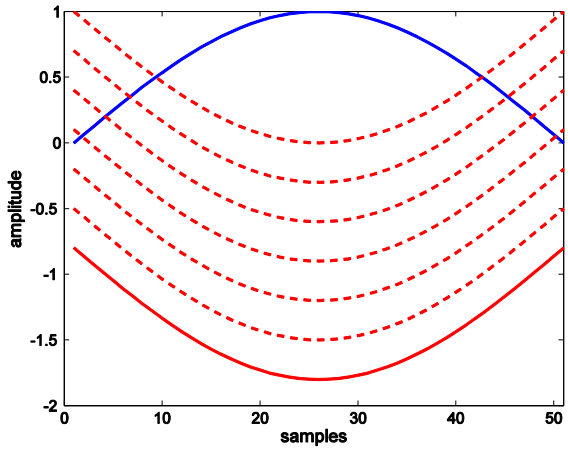


Figure 20: ConstFreqConstTimeDiffRange: Contours with opposing shape (rise-fall vs. fall-rise) are separated by increasing height differences.

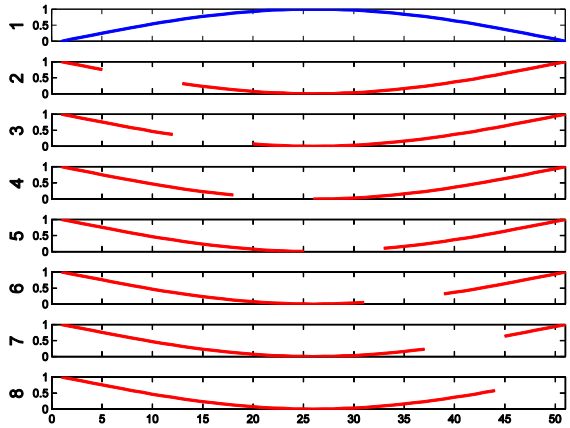


Figure 21: ConstFreqOneGap: A rise-fall contour (blue) is contrasted with fall-rise contours (red). In the red contours, a gap is introduced at varying phase.

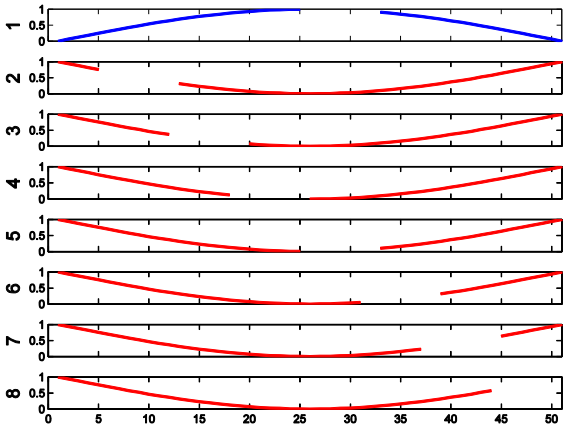


Figure 22: ConstFreqTwoGap: In both contours (blue and red), a gap is introduced. The gaps of contour 5 and of the blue contour coincide.

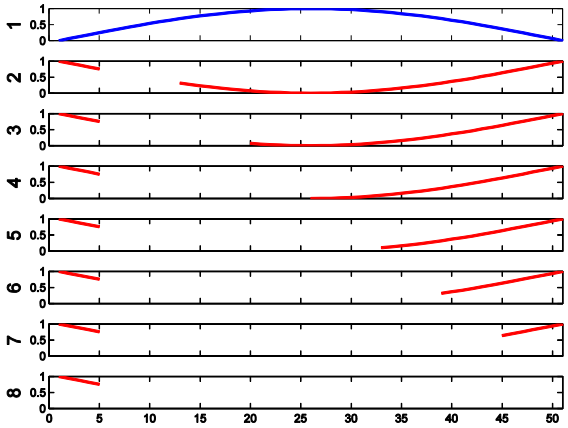


Figure 23: ConstFreqOneGrowingGap: A growing gap is introduced in the inverted contours (red).

5.2.3.3 Window size

One of the basic concepts of the maximum similarity search algorithm requires us to limit the size of the analysis window. For every step of computing the similarity matrix, a window is applied, which averages the similarity of each value pair within that window. If the percentage of non-valid values (missing data) exceeds 20% of the window size, no similarity value is computed for that window. As the two contours may have different durations, and as the window size is bound to be smaller than of both the contours, it is expected that the window size has an influence on the maximum similarity score.

The basic contour in the set of artificial contours comprises 50 samples. When setting the maximum similarity search algorithm, we apply window sizes of 10, 30 and 50 samples, which represent 20, 60 and 100 percent of the basic contour's duration.

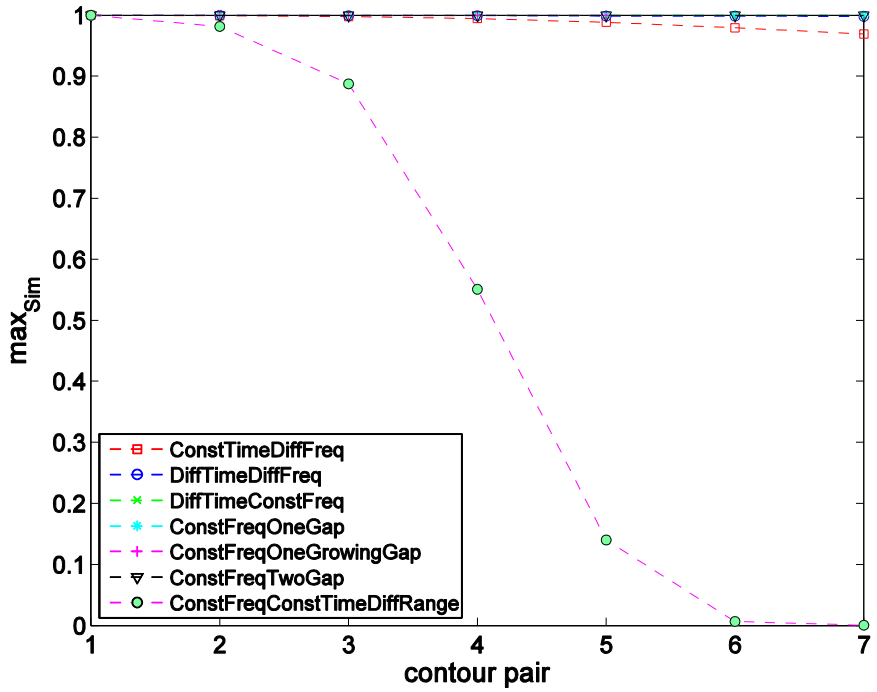


Figure 24: Maximum similarity score max_{Sim} for basically "similar" contour pairs. The window size is 10 samples. There are seven such pairs of contours. For example, the first contour pair consists of contour 1 from Figure 11 (blue contour) paired with contour 2 from Figure 11 (red contour); the second contour pair consists of contour 1 (blue) paired with contour 3 (red) and so on, up to the 7th contour pair that consists of contour 1 (blue) paired with contour 8 (red).

5.2.3.4 Results

The following figures, starting with Figure 24, illustrate the similarity metric (max_{Sim}) calculated by the maximum similarity search algorithm. The scores are arranged according to two sets of contour types: the first set (basically similar) and the second set (contrasting) and according to the applied window size.

Figure 24 illustrates the condition with the window size of 10 samples (approximately 20 percent of the overall duration) and the set of basically similar contour pairs. All contour pairs are rated very high (close to 1), except for the contour pairs with modulated height (ConstFreqConstTimeDiffRange). The following dependence can be seen: the further the two contours are apart in the frequency range, the lower is max_{Sim}.

In more linguistic terms, the maximum similarity score algorithm rated the contour pairs as being very similar for six out of the seven contour types, independent of:

- Differences due to an increase of the velocity of the F0-movement (ConstTimeDiffFreq; see Figure 10),
- Differences due to an increase of the duration, while the velocity of the F0-movement is kept constant (DiffTimeConstFreq; see Figure 11),
- Differences due to an increase in the duration and a decrease in the velocity of the F0-movement, i.e. the same contour shape is produced, just slower (DiffTimeDiffFreq; see Figure 12),
- Differences due to introduced gaps to simulate voiceless segments (ConstFreqOneGap, ConstFreqTwoGap and ConstFreqOneGrowingGap; see Figure 14, Figure 15 and Figure 16).

But a difference in the F0-range between the contours in the contour pair (ConstFreqConstTimeDiffRange; see Figure 13) resulted in a change of the similarity rating.

The further apart the two contours are in the contour pair, the less similar the pair is rated by the algorithm.

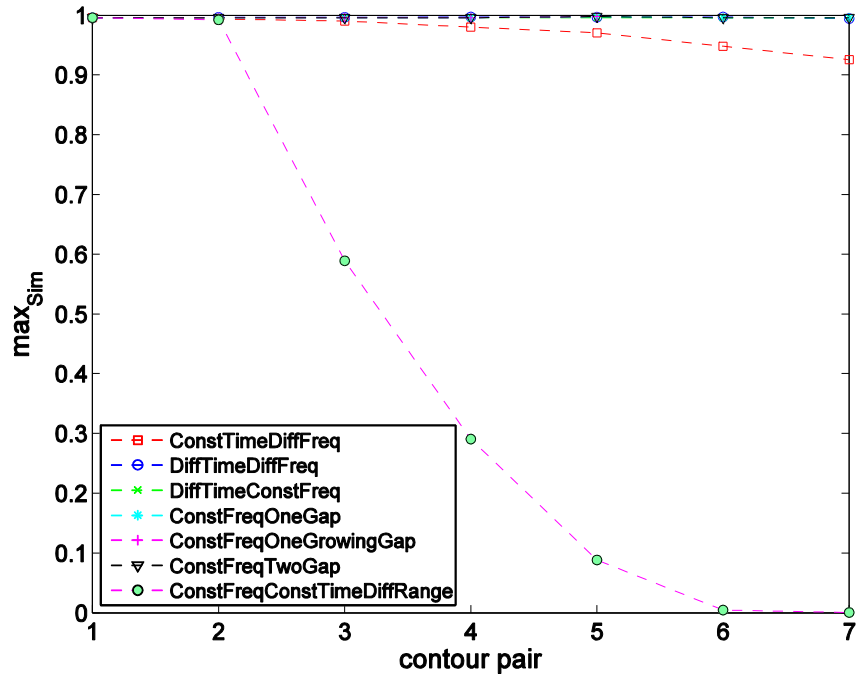


Figure 25: Maximum similarity score max_{Sim} for "inverted" contour pairs with a window size of 10 samples.

This trend also holds for the contours that are designed to be contrastive (inverted) (see Figure 25). Independent of differences in the modulation of the overall shape by manipulating the duration, the velocity of the F0-movement and the introduction of gaps, the algorithm rates the contour pairs as very similar. Again, an exception is the difference in the F0-range that leads to a decrease in the similarity score.

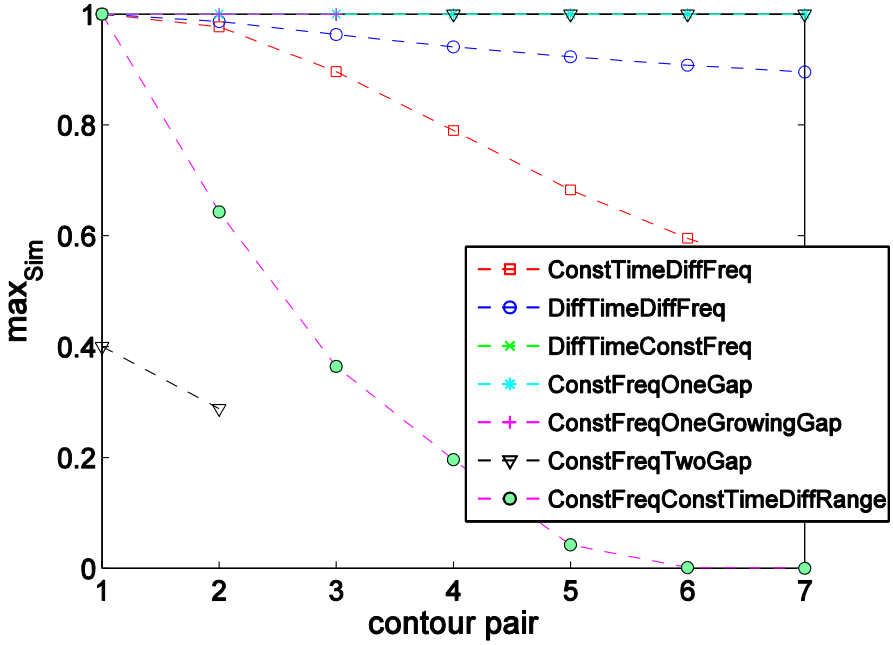


Figure 26: Maximum similarity score max_{Sim} for basically "similar" contour pairs with window size of 30 samples.

Figure 26 and Figure 27 illustrate the condition with the window size of 30 samples (more than half of the samples in the first contour in the contour pair).

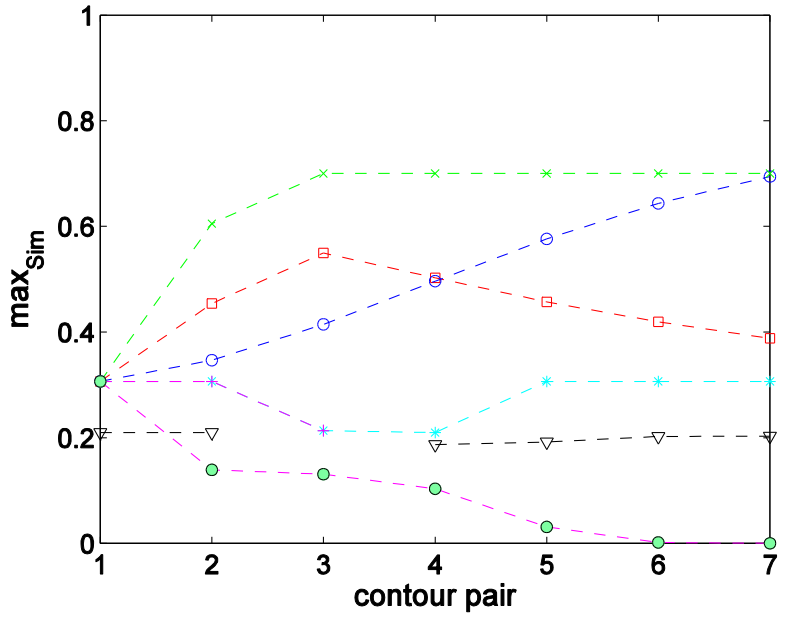


Figure 27: Maximum similarity score max_{Sim} for "inverted" contour pairs with window size of 30 samples. The legend is the same as in the previous figures.

For basically similar contours (Figure 26), the window size of 30 samples caused max_{Sim} to fall due to the change in velocity of the F0-movement and due to F0-range differences. Also the contours with a similar shape but a different duration (DiffTimeDiffFreq) let max_{Sim} decrease with increased stretching. For the contour pairs with gaps, max_{Sim} stays either high (approximately 1), falls down below 0.4 (for ConstFreqTwoGap) or no max_{Sim} is available when no window exists which contains enough valid samples. It seems that max_{Sim} strongly depends on the location and the size of the gaps which cause fluctuations and discontinuities.

In more linguistic terms, a difference in the overall shape due to a change in duration or a change of the velocity of the F0-movement resulted in a decrease of the similarity score. The introduction of single short voiceless segments did not affect the recognition of the overall high similarity of the contour pairs. However, the longer the voiceless segments were, the more likely it is that the algorithm fails to produce any rating of similarity. This is the case when the size of the voiceless segments exceeds the size of the analysis window.

Regarding the inverted contour pairs (Figure 27), the overall similarity score is decreased for all contour types (around 0.3), as expected. But with increasing time differences or frequency differences, max_{Sim} increases (see ConstTimeDiffFreq, DiffTimeDiffFreq and DiffTimeConstFreq). It seems that a larger similarity matrix (more time in either contour) increases the probability to find a matching part. This match does not equal 1 for the overall contrasting contour pairs, because the similarity is averaged along each window. Regarding the contour pairs with gaps in general, max_{Sim} is lower than for contours without gaps. Regarding the pairs with increasing gap-size (ConstFreqOneGrowingGap), once the gap size exceeds the number of samples in the window, no measure for max_{Sim} is available. Contours that are different in height are penalised the most. A height difference of 0.3 reduces max_{Sim} from 0.3 to 0.1. Height differences that are increased beyond this number have little effect. When the height difference is increased up to 1.2, max_{Sim} is almost zero.

In more linguistic terms, the increase of the size of the analysis window from 10 samples to 30 samples has a major effect on the similarity scores obtained from the algorithm. The initial contour pairs drop from a very high similarity score around 1 (see pair 1 in Figure 25) to a relatively low similarity score around 0.3 (see pair 1 in Figure 27), or even lower around 0.2 for the condition where both contours contain a voiceless segment (ConstFreqTwoGap). Increased differences of the overall contour shapes due to increased duration or a change of the F0-movements, however, brought parts of the contours closer, increasing the likelihood that the

algorithm finds parts that are close to each other. Those contour pairs are then rated as more similar.

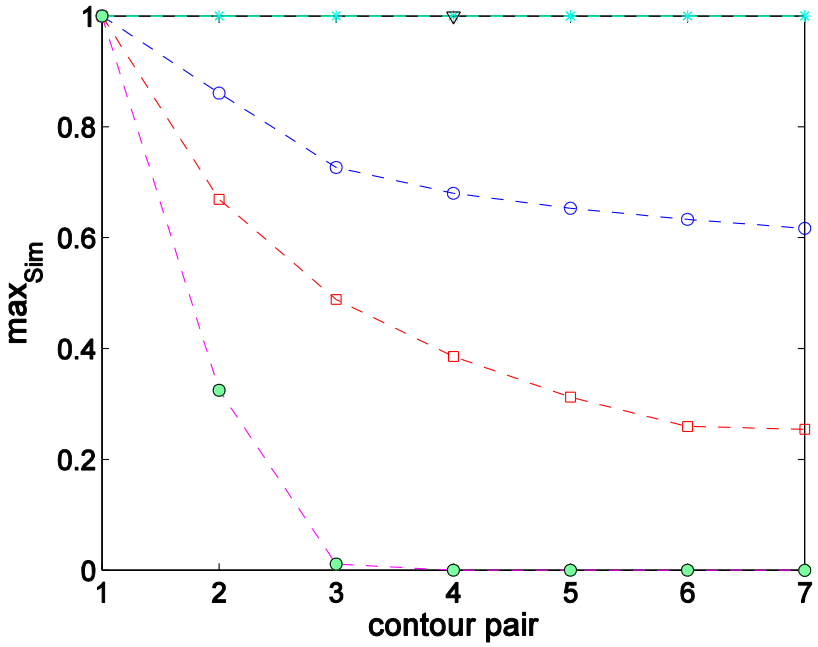


Figure 28: Maximum similarity score max_{Sim} for basically "similar" contour pairs with a window size of 50 samples. The legend is the same as in the previous figures.

Figure 28 and Figure 29 illustrate the condition with the window size of 50 samples (100% of the first contour's duration). For basically similar contours (Figure 28), the comparison using a window size of 50 samples shows that max_{Sim} strongly varies. Differences in height and frequency result in lower max_{Sim} than differences in both time and frequency (DiffTimeDiffFreq). Small gaps have no influence on max_{Sim} which retains the score 1. However, if the size of the gap increases, the similarity measure becomes unavailable. Only one value equalling 1 could be computed for the contour pairs where more than one gap is present in the two contours (ConstFreqTwoGap), as no more than 20% of missing F0-values are allowed by the algorithm.

In more linguistic terms, the window size of 50 samples affects the similarity scores from the algorithm as follows. The F0-range makes the similarity score already drop to a value close to zero at the third contour pair, i.e. a F0-range difference of 0.6 (ConstFreqConstTimeDiffRange; compared with Figure 26). A difference in the overall shape of the contours due to a change of the velocity of the F0-movement (ConstTimeDiffFreq) equally leads to a stronger decrease in the similarity score (compared with Figure 26). Keeping the overall shape constant, but stretching one contour in the pair, results in a stronger decrease of the similarity score. The introduction of short voiceless segments does not affect the recognition of the overall similarity. The more voiceless segments are introduced into the contours, the more the algorithm struggles to produce a similarity score for the contour pair.

Regarding the inverted contour pairs (Figure 29), the overall similarity scores are relatively low, even lower than the scores which result when a 30 samples window size was chosen. An increase in time of the second contour (DiffTimeConstFreq) causes an increase in max_{Sim}. When both, time and frequency are modified (DiffTimeDiffFreq), max_{Sim} remains at the same level. The measures for the other contour types remain relatively constant. An increase in the size of gaps (ConstFreqOneGrowingGap) makes the score of max_{Sim} disappear.

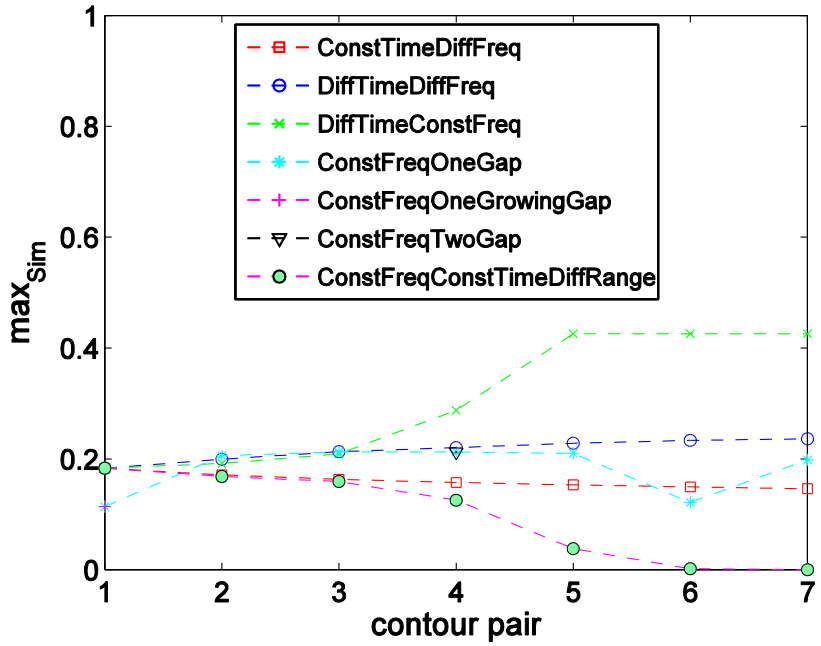


Figure 29: Maximum similarity score max_{Sim} for "inverted" contour pairs with a window size of 50 samples.

In more linguistic terms, an increase of the size of the analysis window from 30 to 50 samples results in an overall decrease of the similarity score from 0.3 to 0.2 for the first contour pair in each modification condition. An increase in similarity due to an increase in duration (DiffTimeConstFreq) is not as strong as with a window size of 30 samples (compared with Figure 27). A change of the overall contour shape due to an increase of the velocity of the F0-movement (ConstTimeDiffFreq) did not lead to an increase in similarity (compared with Figure 27). Keeping the overall contour shape constant while increasing the duration (DiffTimeDiffFreq), did not affect the similarity score. The introduction of a single voiceless segment did not affect the similarity score, but the more voiceless segments are introduced, the more likely it is that the algorithm fails to produce a similarity score.

5.2.3.5 Expectations vs. results

When the contours are the same (e.g. blue contour and first red contour in Figure 10, Figure 11, Figure 12 and Figure 13), the maximum similarity score is expected to be 1, which means perfect match. When either the frequency (Figure 10), the time (Figure 11) or the height (Figure 13) changes, the similarity score is expected to decrease because the contours become more and more dissimilar. This can be attributed to an increased frequency and to an increased height difference, but not to an increased time interval.

When the second contour has the same basic shape, e.g. rise-fall, but is stretched in duration, i.e. time (all contour pairs in Figure 12), the quality score of the contour pair was also expected to be very high. However, the similarity score decreases when the time interval increases. Furthermore, the growing window size reduced the similarity score even more due to the averaging effect of the window.

Regarding contours that have identical shape (e.g. rise-fall), but have different overall height (Figure 13), the similarity score was expected to decrease with growing height difference. The actual similarity scores show this behaviour.

Regarding contour pairs with gaps (in Figure 14 and Figure 15) the score was also expected to be close to 1, as the contours are essentially the same. This high similarity is achieved for some regions, but only as long as the window size does not exceed the amount of valid data for any stretch of the data (cf. to Figure 16). The maximum similarity search then only finds those stretches. For the other stretches no similarity score could be computed.

Regarding all basically contrasting contour pairs (Figure 17 to Figure 23) an overall decrease of the maximum similarity scores was expected. Lower similarity scores were actually reported for these basically dissimilar contour pairs. But the results also show that this strongly depends on the size of the analysis window. It also shows that the similarity score can unfortunately "recover" from dissimilar basic contours when the time of the contours increases beyond the size of the analysis window.

With the "maximum similarity score" metric, as it was used in the paper by Gorisch, Wells and Brown (2012) we were able to show differences in similarity between two social actions. However, the evaluation of the algorithm with artificial contours which systematically modelled the phenomena that we encounter in real recordings (different contour shapes, height differences and sparse F0 readings) exposed shortcomings in the estimation of prosodic similarity.

5.2.4 Discussion of the evaluation

In an ideal case, we would compare two well-defined contours, one from the first speaker's IP and the other from the second speaker's IP. A direct correlation of the two signals could give us an indication of their similarity. But there are two issues to be discussed about the data. First, the two utterances may differ in duration. This means that a direct correlation is not possible. Second, there are regions in the utterance for which no F0 values are available, resulting in gaps in the F0 contours.

The maximum similarity within the similarity matrix is suggested as one measure to indicate the prosodic similarity between the two utterances.

Results show that this maximum similarity score indicates basic differences of the two compared contours, but there are some issues with the maximum similarity score found for a relatively small window in a matrix. For example if there is a fall in F0 at the beginning of utterance A followed by a F0 rise in the same utterance, and the opposite is true for utterance B, these two contours (fall-rise) and (rise-fall) can be considered as being distinct or almost opposite. The similarity matrix is computed across the whole of utterances A and B. Low similarity is achieved where the fall coincides with the rise and high similarity is achieved for the regions where fall coincides with fall and rise coincides with rise, related to the respective contours. If the analysis window is only half as long as utterances A or B, the maximum similarity score within the similarity matrix is found in the high similarity regions. Inevitably, this results in a false positive error, indicating high similarity for a prosodically contrasting contour pair.

The window has been introduced in order to cope with the voiceless regions, however it does not only imply the problem which we have just described, but it also causes discontinuity and unpredictability which are due to the existence of voiceless regions. This is shown by the evaluation results above.

The computation of the similarity matrix is performed on both, utterances of similar duration and on the utterances of different duration. However, our choice of the maximum within that matrix does not reflect this time difference in any way. This is no longer a problem when two utterances of equal duration are compared such as for example "da" and "yeah". Imagine the two utterances are equal in duration. Then, the maximum similarity is the overall similarity. If "da" and "yeah" match each other well, the maximum similarity will be high. But when two utterances are of different duration, for example if we add two more "da" we get "da da da" and "yeah". The short "yeah" may be similar to one of the three "da", but may be dissimilar to the other two. We would treat this as an overall less similar pair than the other pair above. If the duration of that match falls into the size of the analysis window, the maximum similarity search would misleadingly report a high similarity score.

There is one major problem remaining with the method just explained. The longer the analysed turns are, the more likely it is to find a part (window) of the contours, where they match, i.e. have high similarity. But if a match was found in one part of the contours, it does not

automatically mean that the other parts do also match. However, if the turn duration exceeds the window size it implies that false positive results are likely to be derived from that method.

In a compressed form the short "yeah" may still have the same overall prosodic characteristics as the three times longer utterance "da da da". Both utterances are prosodically similar. It may then be argued that the "yeah" is neither fully similar with the first "da", nor with the second or third "da", but only similar with the whole utterance "da da da". The maximum similarity score will therefore be lower than the expected perfect match. In consequence the algorithm does not reflect the overall similarity of the two utterances. Instead, low similarity will misleadingly be reported, presenting a false negative result.

The current linguistic theory requires that two turns of different duration exhibit the same kind of prosodic movement, whatever their durations are. For example a long F0 fall that stretches over a whole utterance (e.g. several syllables) can be similar to an F0 fall that is much shorter and just extends over a single syllable. If for example a single "yeah", which has just the size of the window, the above mentioned method would search for the highest similarity of that window within the prior speaker's long utterance "da da da" and it would not find a match that extends from the start to the end.

In order to find the match over the whole utterance, we require a method that allows for dynamically warping the time (or duration) of the utterances and aligning them accordingly.

Let us restate our general aim: We envisage rendering similarity scores from utterances of equal duration and utterances of different duration comparable.

(a) Imagine a "yeah" which has a similar contour shape as a single "da" which has the same duration.
(b) Imagine a "yeah" which has a similar contour shape as an utterance "da da da" which is three times as long as the "yeah".

Both utterance pairs should be assigned the same similarity score.

The maximum similarity score metric from Gorisch et al. (2012) presented above cannot achieve this. How should we deal with the difference in duration of the utterances? The next section seeks to find an answer to this question by introducing a second metric for measuring prosodic similarity. Nevertheless, the maximum similarity metric has been successfully applied on real data and distinguished action aligning short turns from action non-aligning short turns.

5.3 Accumulative quality score

Regarding the second research question RQ2a asking if alignments and non-alignments differ in prosodic similarity, a further metric is presented here that produces measures of prosodic similarity. It makes use of a technique that compares two complete time series for their similarity from the beginning to the end. It accumulates the pair-wise similarity over the course of the two signals to a final quality score. This represents a further metric that addresses the research question asking for an objective way of measuring prosodic similarity (RQ2b).

5.3.1 Dynamic programming and dynamic time warping

Dynamic programming (DP) is used to solve complex problems by recursively solving subproblems. The solutions of the subproblems are then combined in order to come to an overall solution of the initial complex problem. In our case, the problem is to find the overall similarity of two time series (prosodic information of two utterances). Within DP there exists a technique that suits our purpose.

Dynamic time warping (DTW) is a technique that allows the comparison of two time series with different durations. It aligns the time series according to an optimal time distortion (Rabiner & Juang, 1993). DTW has been used in automatic speech recognition tasks for many years.

Although it has been superseded by statistical modelling techniques in automatic speech recognition, it can also be used to discover re-occurring patterns in two time series, where the beginning and the end of the expected match are not given, as described by Aimetti, Moore, ten Bosch, Räsänen and Laine (2009) and Aimetti, Moore and ten Bosch (2010).

In order to find the best match for signals, DTW has the advantage of allowing the time scale to be dynamically stretched or shrunk. Automatic speech recognition for example employs DTW in order to match a target utterance with a template, even if one was produced faster or slower than the other. (See Holmes and Holmes (2001), Chapter 8.6 or Rabiner and Juang (1993), Chapter 4 for a detailed explanation of that technique.) Here we face a similar problem, as the two utterances that we want to compare may have different durations.

In the case (a) from the section above, no insertions and deletions in the time series of the "yeah" would be needed in order to match the "da" (if both have an overall similar shape). In case (b) however, the "yeah" would need to be stretched in order to match the much longer utterance ("da da da"). (The question arises whether this sort of stretching and shrinking should be kept within limits. For instance, do we want to stretch a 200 ms syllable to match a 2 s sentence? How should we limit the stretching of the time? In DTW these limits are implemented by penalties for each time distortion, as will be explained below.)

The advantage of the dynamic time warping technique over the search for a maximum similarity score is that in order to find the optimal alignment of the two time series, it accumulates the similarity of the two utterances for each point in time which can possibly be passed from the start (first sample of each time series) to the end (last sample of each time series). At each step, the optimal path is chosen and the according similarity score at this point is added to an accumulative quality score. This means that the final point contains the optimal overall quality score of the optimal alignment of the two time series.

Basically, DTW computes a distance matrix. In our case, we compute a similarity matrix instead. The similarity matrix represents the similarity of two utterances at each point. This matrix allows searching for the optimum path with the lowest cost. In our case, this matrix allows looking for the path that maximises a quality score. In order to allow for time distortion, this path may lay outside the diagonal that is drawn from the starting point of both contours to the end point of both contours.

The way in which DTW finds the optimal path through the similarity matrix is done in two steps. In the first step a quality matrix Q is built from the similarity matrix which accumulates the similarity scores in a quality score from the start of the two contours $(1,1)$ up to the end of the two contours (I, J). In the second step the algorithm traces back from the end point to the starting point. There are several possibilities for calculating the quality matrix. One way is described by Aimetti (2011): At each point, beginning at the top left corner of the similarity matrix, the highest similarity score of the three possible predecessors is added to the quality score which has been accumulated up to this point previously and stored in a new quality score at the current point. Given the point (i, j) is reached the three possible predecessors are $s(i-1, j)$, $s(i-1, j-1)$ and $s(i, j-1)$.

$$\phi = max \begin{bmatrix} s(i-1,j) \\ s(i-1,j-1) \\ s(i,j-1) \end{bmatrix} \begin{matrix} \text{horizontal path} \\ \text{diagonal path} \\ \text{vertical path} \end{matrix} \quad (7)$$

If the highest of the three similarity scores is found in point $(i-1, j)$, the new quality score at point (i, j) is stored as the quality score which has accumulated at point $(i-1, j)$ plus the similarity score at the same point $(i-1, j)$. Additionally to the quality matrix, a backtrack matrix is generated which stores at each point the direction to the optimal predecessor. Here, in the backtrack matrix in point (i, j), it is stored that the optimal predecessor was found along the horizontal path. In analogy, if the highest similarity score is found in one of the other points $(i-1, j-1)$ which corresponds to the diagonal path, or $(i, j-1)$ which is the vertical path, the similarity score is added to the accumulative quality score in this point and also stored in point (i, j).

$$q(i,j) = max \begin{bmatrix} q(i-1,j) + s(i-1,j) \\ q(i-1,j-1) + s(i-1,j-1) \\ q(i,j-1) + s(i,j-1) \end{bmatrix} \begin{matrix} if\ \phi = 1 \\ if\ \phi = 2 \\ if\ \phi = 3 \end{matrix} \quad (8)$$

Another way of calculating the quality matrix is described in Sakoe and Chiba (1978). There, the similarity score in point (i,j) is added to the quality scores of each possible predecessor first. Second, the combination value that is the highest is stored in $q(i,j)$.

$$q(i,j) = max \begin{bmatrix} q(i-1,j) + s(i,j) \\ q(i-1,j-1) + s(i,j) \\ q(i,j-1) + s(i,j) \end{bmatrix};\ \phi = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad (9)$$

In order to be able to perform a backtracking step, the value for the direction, where the highest similarity was found (1, 2 or 3 for horizontal, diagonal or vertical) is stored together with the new quality score.

$$\phi(i,j) = \phi \quad (10)$$

The calculation of $q(i,j)$ as described above, favours horizontal and vertical steps over diagonal steps. This is because in order to reach point (i,j) from point $(i-1,j-1)$ over the diagonal path, adds the similarity score found in point (i,j) only once, while reaching point (i,j) over the horizontal + the vertical or over the vertical + the horizontal path adds at each step the similarity of the predecessor. This means that it is theoretically possible that twice as much similarity accumulates over the two steps than over the single (diagonal) step. To avoid this preference, the similarity added for the diagonal path is multiplied by 2.

$$q(i,j) = \begin{bmatrix} q(i-1,j) + s(i,j) * vertical \\ q(i-1,j-1) + 2s(i,j) * diagonal \\ q(i,j-1) + s(i,j) * horizontal \end{bmatrix};\ \phi(i,j) = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad (11)$$

This equalises the preference of vertical and horizontal steps over the diagonal step. In some cases, it might even be wanted to penalise the two steps which cause time distortion of the alignment path by introducing further weighting factors (vertical, diagonal, horizontal). See Sakoe and Chiba (1978) for more detail. (In our study we refrain from doing so, as time distortion should be allowed.)

5.3.1.1 Normalisation for length

When the final point (I,J) is reached, $q(I,J)$ contains the overall optimal quality score which has accumulated over the whole two utterances. This is the optimum quality score as at each point in the matrix the optimal predecessor was determined.

The longer the two utterances are, the higher will be their time variation and the higher will be the accumulative quality score. Therefore the quality score needs to be normalised for the durations of the utterances, in the following also called length. This length-normalised quality score is

$$q_{Lnorm} = \frac{1}{N} q(I,J) \quad (12)$$

where $N = I + J$.

DTW does not make assumptions about temporal compression in speech per se. It can be used as an instrument that measures the distance, or similarity, of two contours and allows temporal distortions in order to match them. As long as no constraints on the time distortion are implemented, two contours of an overall similar shape reach the same score. For example a falling contour over one syllable can be matched with a fall that is distributed over several syllables in a longer IP. But it can also be matched with a fall over one or two syllables followed by low F0 over the rest. However it would match less a contour that has a rise somewhere. Whether this lack of assumptions about the temporal compression in speech is

reasonable can be questioned and it might be argued that some assumptions from the linguistic literature should be applied and implemented in the DTW algorithm, such as penalties for time distortion or boundaries for the start and end of an accent. Here, no penalties are employed and as boundaries, the start and end of the IP are used.

5.3.2 DP algorithm for the current study

The computation of the similarity matrix for the accumulation of the similarity scores is the same as Equation (1) in Section 5.2, as no time window needs to be applied here.

Optionally (but not in the current description) the F0-similarity is weighted by the mean intensity of the two according intensity values at time t.

$$iwsim(x(t), y(t)) = sim(x(t), y(t)) * \alpha(t) \tag{13}$$

where

$$\alpha(t) = \begin{cases} A(t), & if A(t) > 0 \\ 0, & otherwise \end{cases} \tag{14}$$

$$A(t) = \frac{x_i(t) + y_i(t)}{2} + c \tag{15}$$

and

$x_i(t)$ and $y_i(t)$ are simultaneous intensity values to the F0 values $x(t)$ and $y(t)$. The constant c is chosen so that most of the intensity values are above 0 as in Equation (4).

By this way, the similarity at each x, y coordinate is stored in a similarity matrix. This matrix is the basis for the dynamic programming algorithm that should find the best alignment of the two turns.

Our aim is to determine how well the two turns match in terms of their prosody (F0 and intensity). With this measure we try to avoid the problems associated with the maximum similarity measure, which aimed for matching parts *within* the similarity matrix. Such maxima may occur at the beginning or at the end of the one or the other turn, but they cannot indicate the *overall* similarity of the two turns. In contrast to the finding of such local matches, the alignment score needs to be computed from the beginning to the end of the two turns.

For each point in the similarity matrix SM, an accumulative quality score is stored in the quality matrix Q, according to the best local path.

The quality matrix is initialised with the values of the similarity matrix.

$$Q = SM \tag{16}$$

A backtrack matrix is created with the same dimensions, which will store for each coordinate the value of the backtrack which indicates the direction of the previous highest accumulative quality score.

During dynamic programming, the values of the new quality matrix are computed based on the similarity scores and the accumulative quality scores.

$$Q(i,j) = max\begin{Bmatrix} Q(i-1,j-1) + 2 * SM(i,j) * diagonal \\ Q(i-1,j) + SM(i,j) * vertical \\ Q(i,j-1) + SM(i,j) * horizontal \end{Bmatrix}; backtrack(i,j) = \begin{Bmatrix} 1 \\ 2 \\ 3 \end{Bmatrix}, \tag{17}$$

where "diagonal", "vertical" and "horizontal" stand for the penalty values which can be chosen for time distortion (vertical and horizontal paths) or no time distortion (diagonal path). Here, all values are set to 1, which means no penalty for any kind of distortion. There are three possibilities to get from point $(i-1, j-1)$ to point (i, j). The diagonal path is the direct path, which involves only one step. If the first step is vertical to point $(i, j-1)$ or horizontal to point

$(i+1, j)$ first, it involves a second step to reach point (i, j). Since the quality score is accumulated at each step, the vertical+horizontal or horizontal+vertical steps are preferred to the single diagonal step. Therefore, the similarity score along the diagonal path is multiplied by 2 to compensate for this inequality.

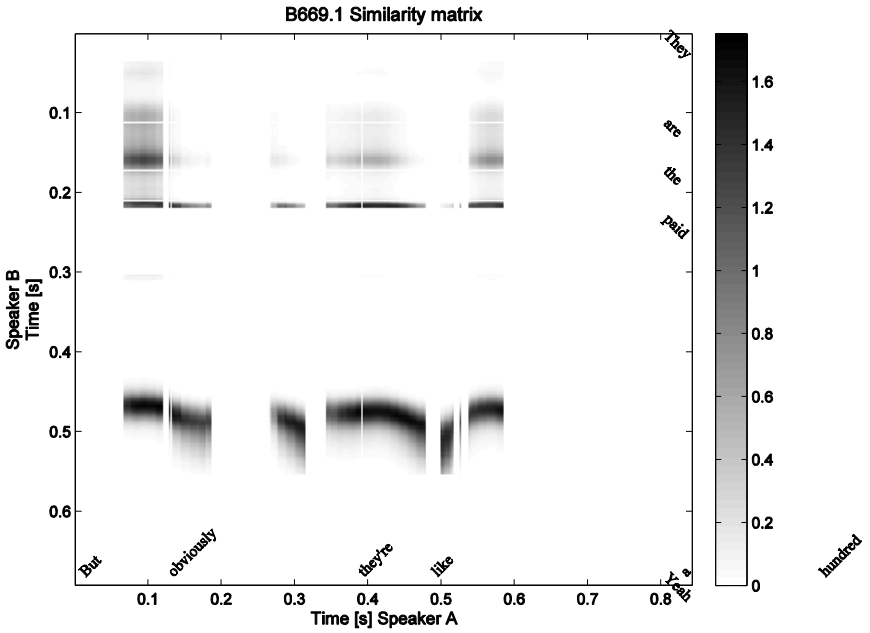


Figure 30: Similarity matrix of the turn-pair "`but obviously they're like`" (speaker A) and "`they are the paid`" (speaker B) based on the DTW algorithm. The pairwise similarity of each F0-value of Speaker A's contour with each F0-value of Speaker B's contour (cf. Figure 6) is indicated on a grey scale. Light grey indicates low similarity and dark grey indicates high similarity.

In the address of the backtrack variable at point (i, j), a number is stored which indicates the path which the local optimal alignment took (diagonal = 1, vertical = 2, horizontal = 3). At the backtracking stage, this information designates the optimum path that leads to the final quality score. The similarity matrix for the current example from Extract 11 "`but obviously they're like`" and "`they are the paid`", is displayed in Figure 30.

The quality matrix with the best alignment path is displayed in Figure 31.

Overlaying the alignment path to the similarity matrix shows the dynamic time warping of the utterances, as is illustrated in Figure 32.

The alignment path is not diagonal, which indicates that the time was distorted in order to match the underlying F0 contours. In the current example, more than half of the target IP (utterance 1) of speaker B is matched with the beginning of the prior speaker's IP (utterance 2), while the rest of utterance 1 is matched with the relatively long remainder of utterance 2.

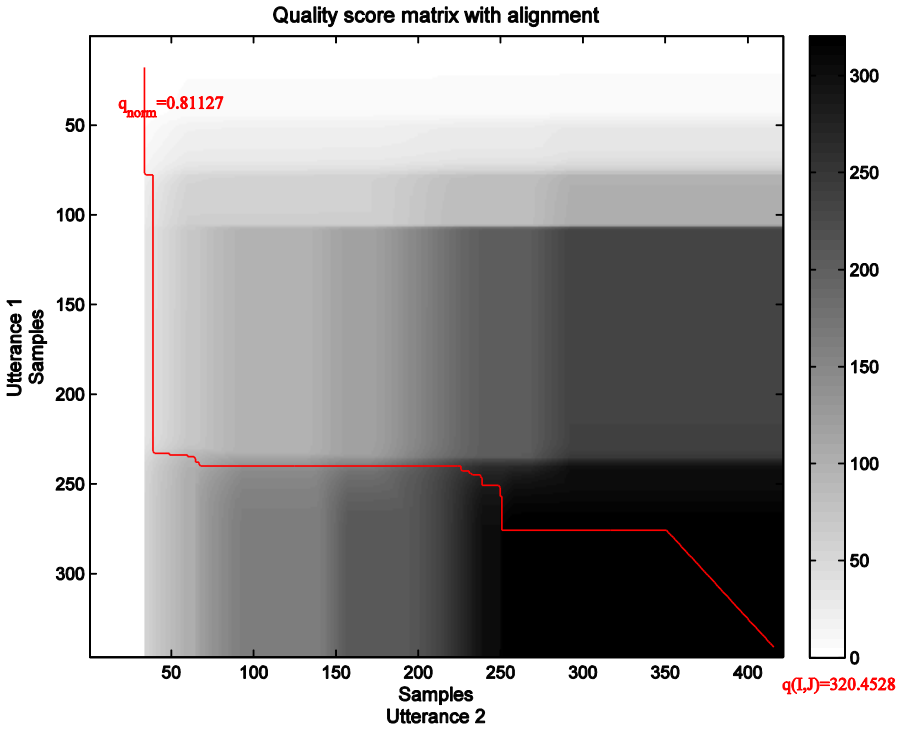


Figure 31: Quality matrix of the turn-pair from Extract 11. The grey scale indicates the accumulation of the similarity scores (cf. Figure 30) from the origin (top left) to the end point q(I,J) (bottom right) of both utterances. The red line shows the optimal alignment path which leads to the highest accumulative quality score (q(I,J)=320.5).

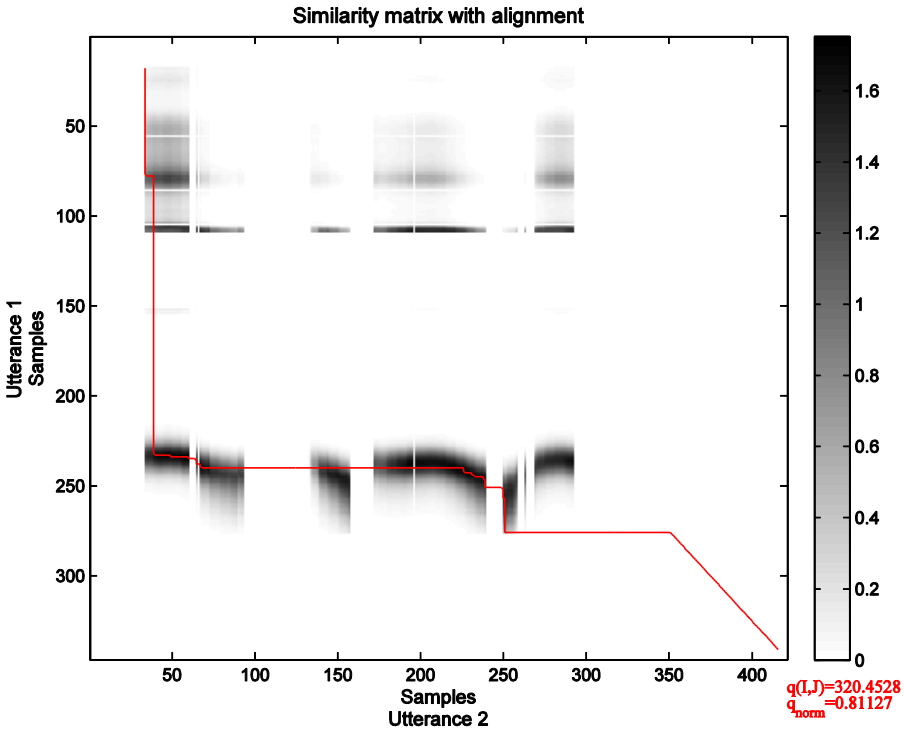


Figure 32: Similarity matrix with optimal alignment for example from Extract 11. At the bottom right corner of the quality score with index (I,J), the accumulative quality score q(I,J) and the normalised quality score q_{norm} are displayed. The alignment path deviates from the diagonal, indicating time distortion. For the optimal matching of the two F0-contours (cf. Figure 6), the beginning and most of utterance 1 (from Speaker B) is matched with the beginning of utterance 2 (from Speaker A), while the rest of utterance 1 is matched with most of utterance 2.

5.3.2.1 Normalisation of the overall quality score

When the similarity scores have been accumulated up to the end of the matrix, the point (I, J) contains the final accumulative quality score ($Q(I, J)$) for the optimal alignment of the two entire utterances.

Normally, the longer the two utterances are, the higher will be the accumulative quality scores at the end. To make quality scores comparable between utterance pairs of different duration, it has to be normalized. The symmetric step pattern used, where the similarity score of the diagonal path is doubled, requires us to divide the final score by the additional duration of the two utterances.

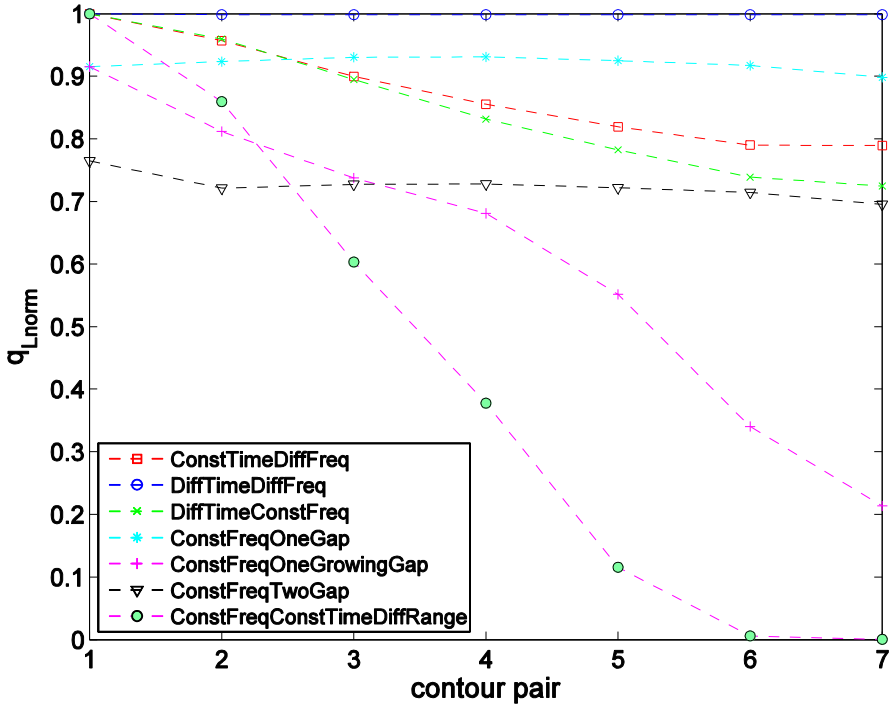


Figure 33: Quality scores normalised for length (q_{Lnorm}) for comparison of basically "similar" contour types.

$$q_{Lnorm} = \frac{1}{N} Q(I, J) \tag{18}$$

where $N = I + J$.

5.3.2.2 Evaluation with length-normalised quality score

For all artificial contour types in Section 5.2.3 (basically “similar” contours a – g) and the contrasting “inverted” contours, the first contour (blue) is compared with the seven other contours (red) of the same type.

We apply the same requirements to the accumulative quality score algorithm as we applied to the maximum similarity search algorithm. An overall difference between the two contour categories of basically “similar” contours (Figure 33) and contrasting “inverted” contours (Figure 34) should be observable. Contours that have the same shape (e.g. rise-fall) should get a high similarity score, independently of the time distortions. Neither the number of gaps, nor their position or size should decrease the value for similarity. Additionally, the overall height differences of the contour pair should be reflected in a decrease of the similarity score.

Regarding the category of basically "similar" contours (Figure 33) the following observations can be made:

The contours with constant time and changing frequency (ConstTimeDiffFreq), show the overall similarity score to decrease from 1 (same contours) to 0.8 (doubled frequency).

The contours with constant frequency and changing time (DiffTimeConstFreq), similarly show the overall similarity score to decrease from 1 (same contours) to below 0.8 (doubled time).

If both time and frequency are distorted simultaneously (DiffTimeDiffFreq), the overall similarity score remains close to 1.

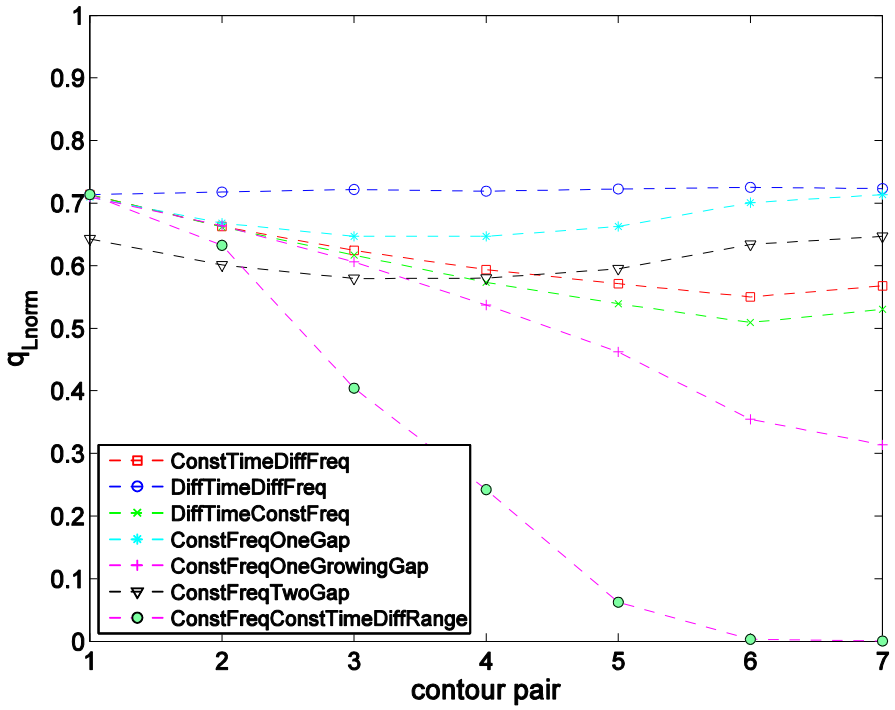


Figure 34: Quality scores normalised for length (q_{Lnorm}) for comparison of the "inverted" contour types.

If the compared contours are separated by their height (ConstFreqConstTimeDiffRange), the overall similarity score decreases steadily and reaches zero at a difference of 1.5. (Regarding the real F0 contours this value would be measured in standard deviations because the contours are normalised for the median and for the standard deviation of the individual speakers).

Regarding the contours with one gap in the second contour (ConstFreqOneGap), the overall similarity score shifts by almost -0.1. Regarding the contours with one gap in each contour (ConstFreqTwoGap), this shift is doubled.

With respect to the contours with an increasing gap in the second contour (ConstFreqOneGrowingGap), the overall similarity decreases constantly and steeper than the changes in frequency or time (alone), but not as steep as the changes in height.

Regarding the category of contrasting "inverted" contours (Figure 34) the following observations are made: The overall similarity decreases to approximately 0.7. The overall relationships between the contour types are similar to the category of basically "similar" contours.

Discussion

With respect to the contours with modified time, frequency and height, the quality score that is normalised for the duration of the contours showed the following results which were expected: the overall similar contours are assigned a high q_{Lnorm}, while the contrasting contours and the contours with increasingly modified parameters (time, frequency, height) show decreased q_{Lnorm} scores.

However, the introduced gaps have a strong influence on the q_{Lnorm} score. A single gap shifts the score down by 0.5, while an additional second gap in the second contour shifts the score down twice as much. It suggests, the longer a gap interval becomes, the more pronounced is that shift. This is the opposite of what we require from the algorithm.

5.3.2.3 Missing data

We explained above how DTW deals with differences in duration of the utterances. DTW allows for time distortions and it tries to optimise the overall alignment of the two utterances. But it is still unknown how the DTW algorithm behaves when it faces missing data, which here comes from the voiceless regions where no F0 information is available. On its way from the origin (i, j) to the end (I, J), the algorithm is forced to pass the missing data area at some point. Since the algorithm always searches for the optimum path, it should also be able to find a path through the missing data region.

Normalisation for missing data

It was described earlier that for the unconstrained DTW algorithm (which does not apply any slope constraint), the quality score is divided by the sum of the lengths of the two utterances $(I + J)$ to get the normalised quality score. But for utterance pairs with a high amount of missing data, the possibility to accumulate a high quality score is lower than for utterance pairs with a low amount of missing data. This means that it is also necessary to normalise the final quality score for the amount of missing (or available) data in the two utterances.

One approach for normalising the final quality score could use the actual alignment path together with the amount of steps through valid (non-gap) regions.

The quality score normalised for missing data would then be:

$$q_{norm} = \frac{L_a}{L_v} q_{Lnorm} \tag{19}$$

where L_a is the length of the alignment, L_v is the amount of steps through valid regions and q_{Lnorm} is the quality score that has previously been normalised for the alignment length.

Hence:

$$q_{norm} = \frac{1}{L_v} Q(I, J) \tag{20}$$

At first glance q_{norm} seems to be dependent on the individual optimal alignment path. But every possible alignment path through the matrix has always the same length $(I + J)$. Every path, e.g. from point $(i - 1, j - 1)$ to point (i, j) has the same step length, as the diagonal paths are multiplied by two $(L_a = N)$. Also the length of "valid" steps will be the same for every possible path. Since no similarity scores can be accumulated in the missing data region, the optimum path will keep the amount of steps through these "non-valid" regions to a minimum. This minimum is the summed missing values in both underlying contours. Its inverse, i.e. the sum of all valid values in both underlying contours, is then L_v.

5.3.2.4 Evaluation with the quality score normalised for missing data

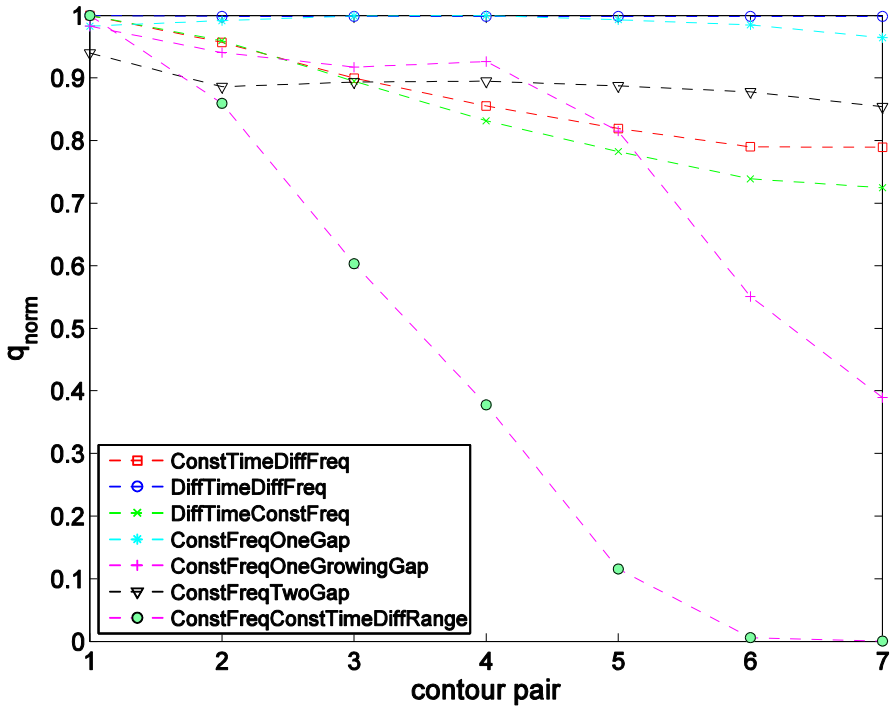


Figure 35: Quality scores normalised for missing data (q_{norm}) compared with basically "similar" contour types.

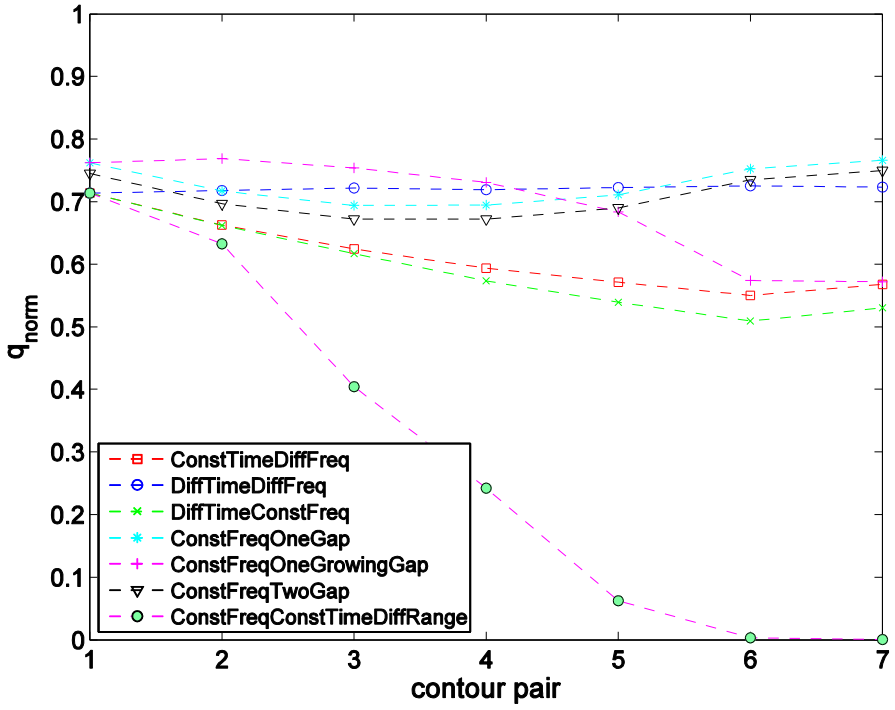


Figure 36: Quality scores normalised for missing data (q_{norm}) compared with the "inverted" contour types.

The quality score that is normalised for missing data is evaluated with the same artificial contours and with the same requirements as the previously discussed similarity metrics. It is predicted that the quality score which is normalised for missing data (q_{norm}) will be to a lesser degree affected by the amount, length and position of gaps than the quality score which was only normalised for the overall length of the two contours (q_{Lnorm}). The distinction should be

maintained between contours having overall similar shape / height and contours having inverse shape and large height differences. Figure 35 illustrates the normalised quality scores for the set of contours that are outlined as basically "similar" contours. Figure 36 illustrates the normalised quality score for the set of contours that are basically "inverted" contours.

Regarding both sets of contours, the similarity scores (q_{norm}) for the contour pairs without gaps remain unchanged in relation to q_{Lnorm}. Regarding the other contour pairs with gaps, the similarity scores seem to have improved in the required direction. Contours with one gap and two gaps have almost identical similarity scores for both sets of contours. When one of the contours has a growing gap, the image between the two contour sets is different. Regarding the basically "similar" contours, an increase in gap size causes a decrease of overall similarity like the decrease due to time or frequency modifications. If more than half of the contour contains non-valid data, the decrease of similarity is even stronger. Regarding the "inverted" contour set, a growing gap has less effect on the similarity score than all the other parameter changes. If more than half of the contour contains non-valid data, the similarity score decreases further and is even stronger than the decrease of similarity scores due to manipulation of time or frequency.

In summary, the modified quality score normalisation that takes missing data into account seems to be an improvement compared with the two previously suggested measures for analysing the similarity of two utterances (one being the maximum similarity search, and the other being the accumulated quality score which was only normalised for the overall duration of the two utterances).

There is still a risk that contours with similar shape will achieve a low similarity score, if these similar shapes have been produced at different height levels. It remains difficult to distinguish between a contour that is just different in shape and a contour which is different in height. Contours which are opposite in shape (i.e. "inverted") cannot achieve high similarity scores, even if they are produced at the same height.

5.3.3 Discussion

If we look at the acoustic-prosodic comparison of two utterances, two basic metrics for a similarity score have been suggested. The first algorithm, from Gorisch, Wells and Brown (2012) correlates the F0 contours and searches the maximum similarity score in the respective calculated similarity matrix. On the one hand, the algorithm provides similarity measures that can distinguish action aligning and action non-aligning turns as has been demonstrated by Gorisch et al. (2012). This is an original contribution to phonetics of conversation research, as it demonstrated that interactional categories that are organised sequentially – not only in the turn organisation, but also in the phonetic organisation – have correlates in the acoustics. These acoustic correlates were analysed automatically by using this algorithm to measure prosodic similarity. However, although the methodology is novel, it has been demonstrated here that there are some limitations of the similarity metric used by Gorisch et al. (2012). It has been shown in an evaluation with artificial contours in Section 5.2.3 that their metric strongly depends on the overall durations of the analysed contours and is influenced by the size of the window which has to be applied. Accordingly, a new algorithm was proposed which is based on the dynamic time warping technique and accumulates the similarity scores of a similarity matrix, for which no window needs to be applied. The overall similarity score is normalised according to the durations of the contours and according to the amount of missing data.

Although the second technique has shown some improvement over the first, the acoustic analysis also depends on the initial normalisation of the F0 contours (see Section 5.1.1). Both techniques show strong dependence on the differences in the height of the contours. If the mean (or median) of a speaker's range is not identified correctly, the error would be maintained along all subsequent steps of the analysis. If two F0 contours, which are perceived as being on a similar F0 level (even if the female and male voices are very far apart from each other), are not brought to a similar level by virtue of normalisation, the comparison will declare the two contours wrongly as distinct. This error can also occur in the other direction: If two F0 contours,

which are perceived as being on a different F0 level, but which are brought (falsely) to a similar level, the comparison will declare the two contours wrongly as similar.

Both metrics, the "maximum similarity score" and the "accumulative quality score" have their advantages and disadvantages for measuring the overall similarity between two contours that can contain missing data. Variation in time and frequency should be allowed without decreasing the similarity measure, if the overall shape, e.g. rise-fall is matched with another similar shape, which only extends over a longer stretch of time. This seems to be better and more consistently modelled by the accumulative quality score in general and by the accumulative quality score which is normalised for missing data in particular.

Given that the initial normalisation of all F0 and intensity values reflects the true speaker's ranges, the suggested metrics constitute measures that can be applied to real data.

The (real) data from the AMI corpus that we use here contains many unvoiced regions, resulting in large proportions of missing data. Our data is therefore quite different to the data Hermes (1998a, 1998b) or Rilliard et al. (2011) had available, and for which linear interpolation of missing data was an option.

5.4 Results on real data

It is hypothesised that aligning and non-aligning actions are reflected in their prosodic characteristics, especially in their sequential use of prosody. This means that aligning and non-aligning turns can be distinguished according to their prosodic match with the immediately prior turn.

The two similarity metrics are now applied to the two adjacent turns of the speakers involved in the AMI meetings. First, we assess the results for the maximum similarity search and second we assess the results for the accumulative quality score. In the article by Gorisch, Wells and Brown (2012), it has already been demonstrated that the two interactional categories (alignment and non-alignment) depend on the prosodic similarity measured with the maximum similarity search algorithm. However, as we have argued above, the underlying dataset was relatively small. In the current study, the collection of instances of adjacent turn pairs is substantially increased from 177 to 908. This makes it possible to assess whether the findings by Gorisch et al. (2012) can be confirmed with a larger database.

In the three meetings (B, C and D) 912 adjacent turn pairs have been identified and interactionally categorised. 4 of them were excluded because the time span between the two turns exceeded 3 seconds, which is considered to exceed the window of subjective presence (Pöppel, 2009). 908 instances (560 alignments and 348 non-alignments) are left for the acoustic analyses. Both metrics, the "maximum similarity search" and the "accumulative quality score" are analysed and compared.

5.4.1 Maximum Similarity search

The maximum similarity search algorithm uses a window for which a certain percentage (80%) of valid F0 values is required to compute the maximum similarity score. 396 contour pairs had only sparse or no F0 information, leaving 512 instances (307 alignments and 205 non-alignments) with sufficient F0 information in the window with the highest similarity. The distribution of the resulting maximum similarity scores (max_{Sim}) according to the interactional categories are shown in Figure 37. Here, only information from the F0 contours is used (no intensity weighting) and the sigma value is 0.3. Both, alignments and non-alignments have a u-shape distribution with many similarity scores close to 0 and close to 1, while the values in the middle are sparse. However, there seems to be a tendency that alignments have more values close to 1 than close to 0, while for non-alignments this seems to be the opposite with more values close to 0 than close to 1.

A similar pattern can be observed for the distribution of max_{Sim} if the algorithm is applied with intensity weighting (same sigma of 0.3), as shown in Figure 38.

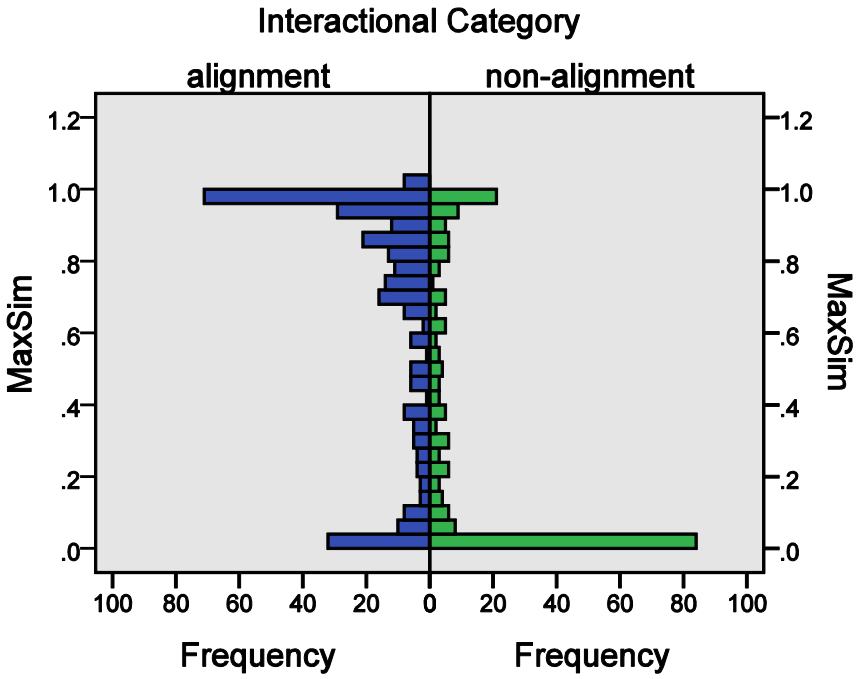


Figure 37: Distribution of maximum similarity scores (max_{Sim}) according to the interactional category (alignment or non-alignment). For the computation of the underlying similarity matrix, only the F0 contour was used and the sigma value was 0.3.

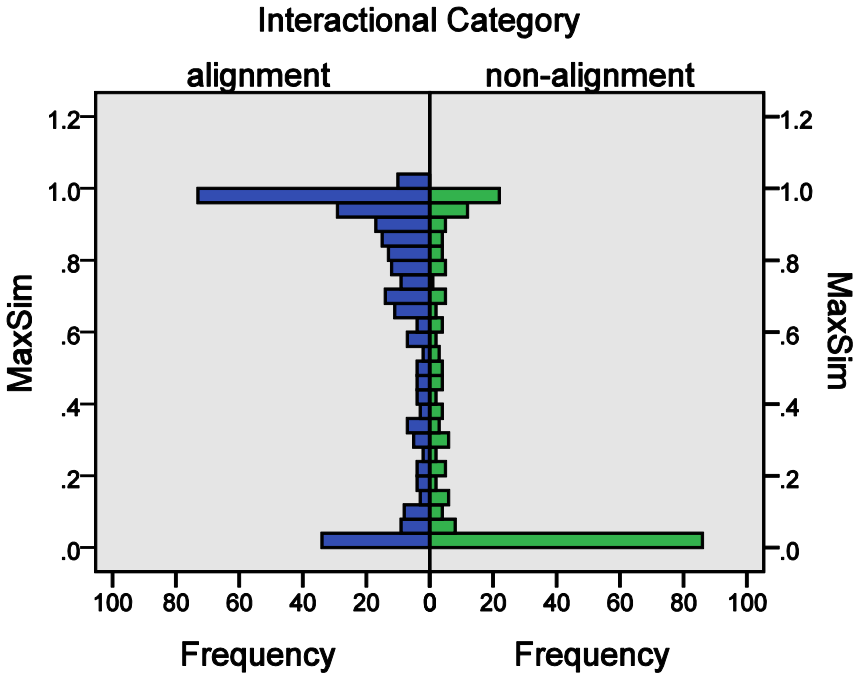


Figure 38: Distribution of maximum similarity scores (max_{Sim}) according to the interactional category (alignment or non-alignment). For the computation of the underlying similarity matrix, the method includes intensity weighting. Sigma is 0.3 as in the case of the illustration in Figure 37.

5.4.1.1 Statistical test

In order to test whether these tendencies of two variables, i.e. the expected values for the similarity score between the two categories, are statistically significant, a t-test is usually employed. But a t-test requires the data to be normally distributed, which is not the case here. Therefore, we abstain from the t-test on the level of the interval scale and constrain the analysis to statements about the central tendency on a lower level, the ordinal scale. An appropriate method is the Mann Whitney U-test, which can cope with skewed distributions and which is therefore "distribution free".

We wish to test whether the medians – the measure of central tendency, which can reasonably be applied on the level of ordinal scales – differ for the two categories. The first step is to arrange all scores on one single ordinal scale and assign to each value a rank on that scale. If one of the two categories (e.g. alignment) had higher similarity scores than the other (e.g. non-alignment) this case would presumably be reflected in a way that the sample of alignments is represented by an overall higher rank than the sample of non-alignments. In other words, in this case we would expect a higher sum over all ranks in the alignment sample than in the non-alignment sample.

For one condition, the significance test is calculated explicitly. All the other conditions are presented in a summarised form.

In the current case, the sum of ranks of the alignments and non-alignments are $T_{align} = 92{,}713$ and $T_{non-align} = 38{,}615$. The mean rank of alignments is 286.12 and the mean rank of non-alignments is 182.63. The higher mean rank for the sample of alignments indicates indeed, that in the alignments sample there are higher rank places. But the question is if we can conclude from the drawn samples whether this is also true for aligning and non-aligning turns in general. First, the two samples are of different size. Second, we can assume that a general fluctuation of rank places takes place.

The U-value helps us to evaluate the circumstances and the sum of ranks which allow us to draw conclusions about the differences between the categories. For each alignment we calculate, how many non-alignments have a higher rank. The U-value is the sum of these non-alignments with higher rank. A simplified equation is

$$U = T_{align} - \frac{N_{align} * (N_{align} + 1)}{2}$$

Where N_{align} is the number of alignments.

For $\alpha = 0.05$ the critical z-value is 1.65. And the critical U-value:

$$U_{crit} = 34{,}173.95$$

Conventionally U is an integer. Therefore the critical U-value is 34,174. Because the empirical U-value (45,435) is higher than the critical U-value we can reject the null-hypothesis in favour of the alternative hypothesis that alignments in general tend to have higher similarity scores than non-alignments.

Here this means that the similarity metric which searches for maximum similarity scores tends to have higher scores for the underlying parameters: only F0 contours (no intensity weighting) and sigma = 0.3.

Regarding the current case, the results indicate mean ranks of alignments of 302.00 and for non-alignments 188.37. The z-score is 8.537 with an according p-value below 0.000 indicating a highly significant difference between the similarity scores of alignments and non-alignments (two-tailed).

Regarding the intensity weighted F0 similarity scores (cf. Figure 38), by keeping the sigma value constant, 280 alignments can then be compared with 204 non-alignments. The mean rank of the alignments is 286.12 and the mean rank of non-alignments is 182.63. The z-score is 8.075 which relates with a p-value below 0.001.

All p-values and z-scores for parameter combinations of sigma and the analysis method (with or without intensity weighting) are summarised in Table 7.

All p-values are far below 0.001, as no precise value can be computed above a z-score of 5. This means that the two categories alignment and non-alignment are statistically significantly different for all combinations of parameters.

The z-scores, which are converted into p-values, are all close to 8 and increase with the sigma value. In the pairs of F0-only and intensity weighted similarity scores (Int-F0), the z-score is slightly higher if the F0-similarity was not weighted by intensity. A test of repeated measures should help to clarify the difference between the two analysis methods (F0 only and intensity weighted F0 similarity).

Table 7: Summary statistics for different sigma values (0.1 to 0.4) and analysis methods (F0-only or Int-F0).

	sigma							
	0.1		0.2		0.3		0.4	
	F0-only	Int-F0	F0-only	Int-F0	F0-only	Int-F0	F0-only	Int-F0
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
z-score	8.299	8.294	8.492	8.422	8.537	8.466	8.614	8.463

5.4.1.2 Discussion

Our hypothesis states that aligning turns have a higher prosodic similarity to the preceding turn than non-aligning turns. The similarity metric that searches for the maximum similarity score within a similarity matrix of the two turns is not normally distributed for both categories. This is the case for all parameter combinations, for the sigma values and for the two analysis methods (F0-only and intensity weighted similarity) which have an influence on the underlying similarity matrix. The distributions suggest that a higher maximum similarity score is achieved on average for alignments than for non-alignments. Due to the non-normal distribution, a Mann-Whitney U-test was applied to evaluate whether the two categories do indeed differ significantly. The first expectation that the aligning turns have scored higher maximum similarity scores than non-alignments is approved throughout the parameter combinations.

The second expectation that intensity weighting of the F0-similarity would increase this difference between the similarity scores could not be proven. However no significance test was used. The observed trend contradicts this expectation, as the z-score decreases when intensity weighting is applied compared to when it is not applied.

5.4.2 Normalised accumulative quality score

Dynamic time warping could be applied to 871 of the 908 pairs of domain and target IP. For 37 pairs no F0 information was available either in one of the two utterances or in both. At the interactional categorisation stage, 533 of the target IPs were classified as alignments and 338 were classified as non-alignments. The distributions of similarity scores (md_{norm}) according to the dynamic time warping algorithm are displayed in Figure 39. There, the method for calculating the underlying similarity matrix refers to the F0 contours only, the sigma value being 0.3. On the one hand the distribution of similarity scores for alignments is skewed to higher values (close to 1), except for 40 instances which fall into the 0 bin. On the other hand, the non-alignments in the 0 bin have a relatively high proportion (more than 80 instances), while the other values are evenly distributed. On average, alignments seem to have a higher similarity score than non-alignments.

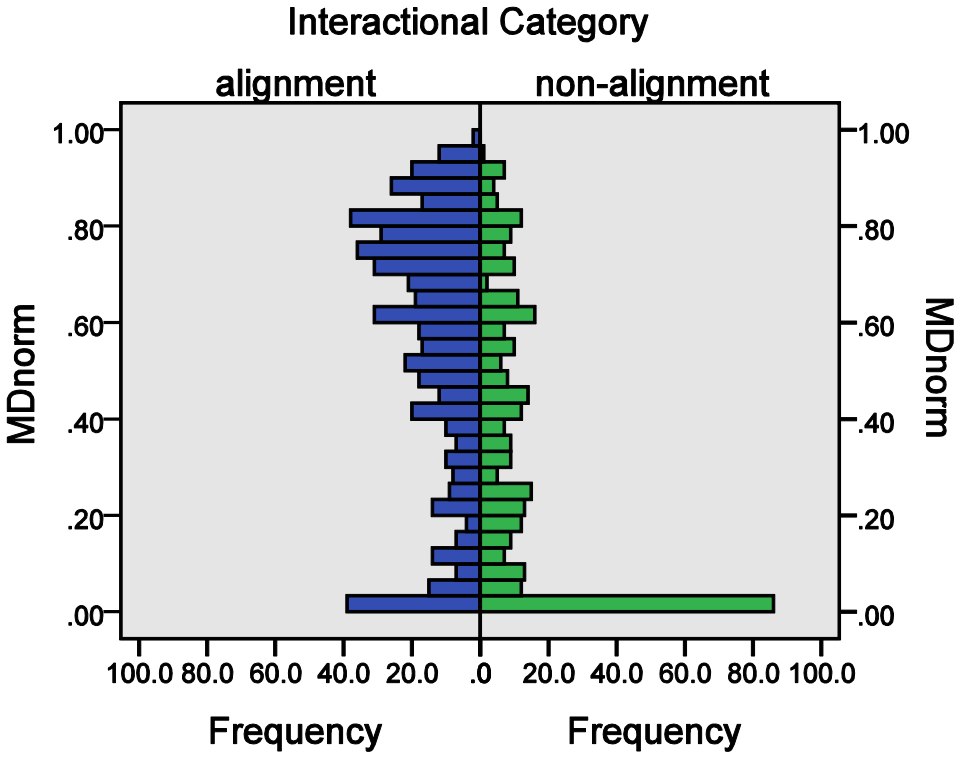


Figure 39: Distribution of quality scores normalised for missing data (md_{norm}) according to the interactional category. The underlying similarity matrix was computed with the F0 contour only and with a sigma value of 0.3.

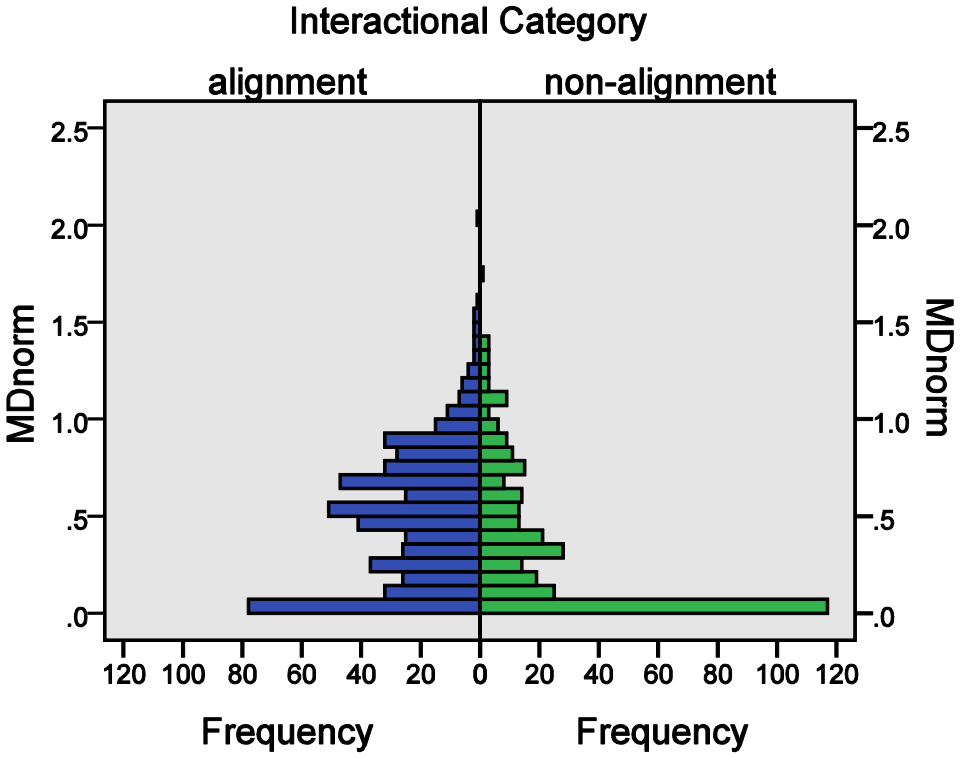


Figure 40: Distribution of md_{norm} according to the interactional category. The method to compute the similarity matrix uses intensity weighting of the F0 similarity. Sigma is 0.3.

The distribution of similarity scores involving intensity weighting at the stage of computing the similarity matrix is shown in Figure 40. This figure is completely different from Figure 39. Intensity weighting made it possible to achieve similarity scores above 1. It is difficult to make predictions for alignments and non-alignments in the intensity weighting condition. But there seems to be the same amount of alignments with scores between 0 and 0.5 and scores between 0.5 and 1.0, while the non-alignments are more skewed to the low values.

For both analysis methods and different sigma values, Mann Whitney U tests (Mann & Whitney, 1947) are conducted. The results are summarised in Table 8. The statistical test needed to be a non-parametric test because the data is not normally distributed.

The null-hypothesis was chosen so that the similarity scores between alignments and non-alignments are equal. An alternative hypothesis states that the similarity scores are different.

With the method of using F0 information only and applying a sigma value of 0.3 in the calculation of the similarity matrix, the similarity scores (md_{norm}) of alignments were found to be higher than the similarity scores of non-alignments, as indicated by the finding that the p values are below 0.001 and the z-scores are above 9. On the average, alignments have a higher similarity score than non-alignments. When varying the sigma values from 0.1 to 0.3, the z-scores increase to a value above 10; when sigma reaches 0.4 the z-score starts to decrease.

Using the method that applies intensity weighting to the computation of the similarity scores (Int-F0) the differences between alignments and non-alignments are also significant, but the z-scores are lower than in the case when the analysis method is applied which is only based on F0. Although the statistical test indicates a significant difference between the two interactional categories according to the similarity metric, the distribution of the data does not make this trend very obvious. There are many alignments that have very low similarity scores, but this is the case for non-alignments too.

Table 8: Summary statistics for different sigma values (0.1 to 0.4) and analysis methods (F0-only or Int-F0).

	sigma							
	0.1		0.2		0.3		0.4	
	F0-only	Int-F0	F0-only	Int-F0	F0-only	Int-F0	F0-only	Int-F0
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
z-score	9.522	7.321	9.934	7.193	10.057	6.742	9.846	6.179

5.4.3 ROC curve analysis

An effective method of evaluating the quality or performance of a diagnostic test is the receiver operating characteristic (ROC) curve. It is mainly used in medical studies (Park, Goo, & Jo, 2004), but can equally be applied to the metrics and methods used in our study, as the categories alignment and non-alignment are binary classes and the similarity scores are measured on a continuous scale. According to Fawcett (2006), ROC graphs are "especially useful for domains with skewed class distributions", which is the case here.

The basic concept of ROC analysis is a mapping from instances to predicted classes, for example to the classes positive or negative. An attempt to classify the collection of second turns into the classes positive (alignment) or negative (non-alignment) will result in four possible outcomes. If the prediction of an outcome is alignment and the actual value is also alignment, it is a true positive; however if the actual value is non-alignment, it is a false positive (type I error). Conversely, if the prediction is non-alignment and the actual value is also non-alignment, it is a true negative; if the prediction is non-alignment, but the actual value is alignment, it is a false negative (type II error).

Here, a binary decision is made in the class prediction between alignment and non-alignment by applying a threshold to the similarity score. If the score is above the threshold score, the

decision is alignment. If the score is less than or equal to the threshold score, the decision is non-alignment. Thereby some correct decisions are made, but inevitably some instances are classified wrongly and fall into the boxes of false positives or false negatives. The latter indicate the errors, or confusion, between the classes.

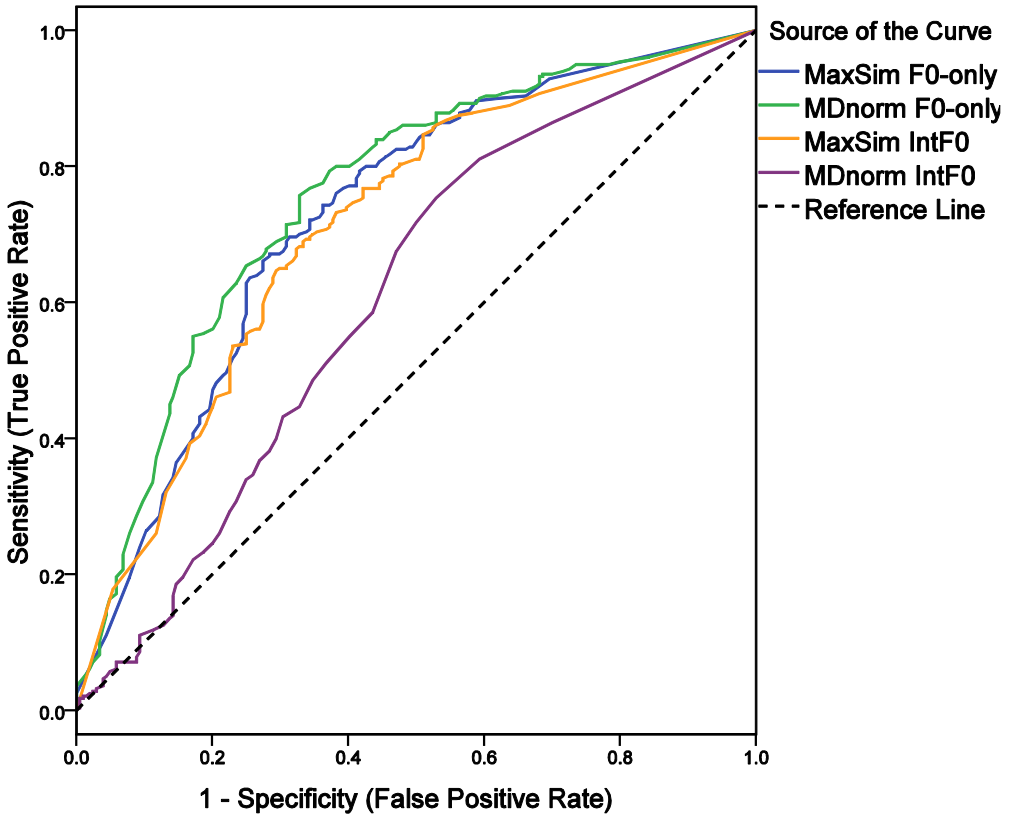


Figure 41: ROC curves for the four combinations of the different algorithms (max_{Sim} and md_{norm} (DTW)) and the different methods (F0-only and IntF0). The curve for max_{Sim} F0-only (blue line) uses the similarity values from the Maximum Similarity Search algorithm and the F0-only method. The curve for md_{norm} F0-only (green line) uses the similarity values from the Accumulative quality score algorithm and the F0-only method. Analogue for the intensity weighted (IntF0) similarity scores and the two metrics. The reference line (dashed) indicates the chance level.

According to these decisions, the ROC curve plots the true positive rate (TPR) vs. the false positive rate (FPR) at different threshold settings. The true positive rate is the fraction of true positives out of the positives (here, alignment).

$$TPR \approx \frac{Positives\ correctly\ classified}{Total\ positives}$$

The false positive rate is the number of false positives (alignments) out of the negatives (non-alignments).

$$FPR \approx \frac{Negatives\ incorrectly\ classified}{Total\ negatives}$$

The area under the curve (AUC) can be used to quantify how well the measurements separate the collection being tested into alignments and non-alignments. It may range between 0.5 and 1, where 0.5 corresponds to the area under the reference line indicating chance level and 1 corresponds to a perfect classification.

5.4.3.1 Results

Figure 41 shows the ROC curves for the two metrics and the two analysis methods. 280 alignments and 204 non-alignments from the overall 909 instances are available for the analysis (for 425 instances in either group of metric or method no similarity score was available).

Table 9 summarises the statistics on the area under the ROC curve for our collection. All measurements (metrics and analysis methods) classify the alignments and non-alignments above chance level (all p-values are below 0.001). The combination of the accumulative quality score (md_{norm}) with the F0-only analysis method reaches the highest area with 0.756. However, it is not possible to tell whether this combination is significantly better than the second or third placed combinations (max_{Sim} + F0-only and md_{norm} + IntF0), as the confidence intervals overlap. The only combination with a significantly smaller area is the md_{norm} + IntF0.

Table 9: Statistics on the area under the curve (AUC) of the ROC curves in Figure 41.

Variables	Area	Std. Error	p-value	95% Confidence Interval	
				Lower Bound	Upper Bound
max_{Sim} F0-only	0.727	0.024	0.000	0.681	0.774
md_{norm} F0-only	0.756	0.023	0.000	0.711	0.800
max_{Sim} IntF0	0.714	0.027	0.000	0.667	0.761
md_{norm} IntF0	0.608	0.027	0.000	0.556	0.660

According to the scale in Table 10 the first three measurements are fairly good in distinguishing alignments from non-alignments, while the fourth measurement is only poor.

Table 10: Quality scale for ranking the area under the ROC curve from Tape (2003).

AUC (area under ROC curve)	Quality
0.90-1.00	excellent (A)
0.80-0.90	good (B)
0.70-0.80	fair (C)
0.60-0.70	poor (D)
0.50-0.60	fail (F)

5.5 General Discussion about similarity metrics

The maximum similarity search algorithm has initially been presented by Gorisch, Wells and Brown (2012) as a similarity metric that indicates the similarity of two utterances according to their prosodic characteristics, namely F0 and intensity. This algorithm was applied in their study to a collection consisting of 177 instances (namely short inserted turns). The results indicated a significant difference between the two groups. This seems to be the first empirically motivated attempt to measure prosodic similarity of adjacent turns in naturally occurring talk in order to distinguish social actions. It has been applied successfully. However, the number of instances was relatively small and there was a large discrepancy between the amount of alignments (149) and non-alignments (28) and a return to an enlarged dataset was envisaged.

In the current study we have increased the overall number of instances. The collection of alignments and non-alignments has been made more balanced by including adjacent turns which don't represent mere "insertions". Its statistical analysis is therefore more robust. The observations from Gorisch et al. (2012) about alignments having, on average, a higher maximum similarity score than non-alignments can now be confirmed.

The maximum similarity search algorithm offers the possibility to add evidence to the expectation that short turns from a second speaker which are designed as agenda alignments are prosodically more similar to the prior speaker's utterance than non-alignments. However, an

objection may arise regarding the question of whether the maximum similarity search is a good enough measure that can depict the true prosodic characteristics. Searching for a single maximum within a matrix partly ignores the detailed characteristics of the underlying contours and their development over time. This can potentially lead to the difficulty that contours which are basically different, falsely achieve high similarity scores, if by chance two parts of the contours happen to be relatively similar, while the rest of the contours are in contrast with each other (see Section 5.2.3 for a longer discussion of that issue).

An alternative metric was devised which takes the entire contour and its time development into account. This algorithm is based on the dynamic time warping technique. It was expected to overcome the drawbacks of the maximum similarity search algorithm, which has indeed been demonstrated for different sets of artificial contours. Applied to the real data, it was expected that the DTW approach can at least show the differences between alignments and non-alignments, as the maximum similarity search metric has demonstrated, or have even better performance. Indeed, for F0-only, the z-scores are higher when the similarity metric using the DTW algorithm is applied, compared with the maximum similarity search algorithm. However, a further analysis using the area under the ROC curves did not prove a superiority of the one algorithm against the other.

It has further been suggested (cf. Section 2.2.4.2) that weighting the F0 similarity by the average intensity with which the F0 values have been produced is a way to reflect the perception of the F0 as prosodic property. In the view of our hypothesis of prosodic matching such integration which approximates prosody in its holistic form is expected to increase the differences in similarity scores between alignments and non-alignments. Such an increase was observed by Gorisch et al. (2012) when applying the maximum similarity search algorithm, although it was not tested for statistical significance. In the current study, this trend could not be observed for that algorithm. Using the DTW algorithm, the z-scores did even decrease when intensity weighting was applied. This difference was even statistically significant, as the analysis of the area under the ROC curves shows. It is tempting to conclude that the intensity is an irrelevant factor in prosody. But this does not necessarily mean that intensity does not play a major role in the perception of prosodic matching and prosodic contrast. Instead we should conclude that the method of intensity measurement – and possibly also the prosodic similarity measure itself – have room for improvement. For example the issue remains of whether intensity is reliable, given variations in the placement of microphones etc. A level-independent metric or a metric that is less dependent on level might be a future enhancement.

One difference between the two algorithms should however be mentioned which points to their usability. For many instances of the collected adjacent turns, the maximum similarity search algorithm could not compute the prosodic similarity at all, due to lacking F0 information. From 909 examples, 397 were missing in the F0-only condition and 425 in the IntF0 condition. Whereas the accumulative quality score algorithm did produce similarity scores for most of the examples. Only 38 instances were missing because no F0 information was available at all. In a practical implementation of the algorithm, this would give an advantage to the accumulative quality score algorithm.

Summarising the findings according to the research questions on the acoustic part of this study, the following can be said. RQ2a, asking if alignments are produced with prosodic matches and non-alignments are produced with prosodic non-matches, can be confirmed on the basis of the data measured by the similarity metrics. RQ2b, asking how prosodic similarity can be measured objectively, cannot be answered directly. However, the metrics work automatically on the audio recordings, suggesting high objectivity. How far the similarity in terms of prosody is captured is unclear. However, the metrics were tested on artificial data and achieved reasonable results. RQ2c can also be answered only indirectly. The acoustic parameters chosen for this study were F0 and intensity. Both parameters, alone and in combination, made it possible to distinguish social actions according to the alignment category, suggesting that they are responsible for the identification of prosodic matches and non-matches. The normalisation of these parameters according to individual speaker differences is considered to be important as is discussed in the following section.

5.5.1 Summary of problems

The direct analysis of consecutive turns from different speakers seems to be a relatively easy task, but there are many problems which were uncovered during the implementation of this acoustic analysis. They have to do with the individual speaker characteristics, the technical aspects of the recordings and the characteristics of the voice in general.

5.5.1.1 Normalisation of speaker differences

When comparing the artificial contours we have seen that the similarity scores strongly depend on height differences. If the compared F0 values are close to each other, a high similarity is achieved. If the compared F0 values are far apart, a low similarity is achieved. A good method would be crucial that allows to initially normalise the F0 and intensity contours according to the individual speaker's voice characteristics (median and standard deviation). On the one hand, if the two contours are already apart initially, due to faulty normalisation, the similarity score will be low, even if the two utterances are heard to be similar. On the other hand, if the two utterances are heard to be different, a faulty normalisation can bring them close to each other and make them appear to be similar.

The method of normalisation that we applied in the current study uses for each speaker the F0 values over the whole meetings. This decision was influenced by the following reasoning: On the one hand, this type of averaging has the advantage that it spreads the risk of using unrepresentative stretches of talk for the normalisation process. On the other hand, the speaking style and pitch range of a speaker at the beginning of a meeting (that takes for example one hour) may be different to the pitch range at the end of a meeting, or even at another meeting on another day. Averaging over shorter stretches of talk could also have an advantage, as Kousidis et al. (2008) and Kousidis et al. (2009) have shown that speakers show acoustic convergence to the other speakers during the dialogue. This means that a continuous normalisation, according to a smaller time windows (e.g. 60 seconds) might improve the analysis. However, if convergence applies, convergence applies to each speaker simultaneously and the effect of the latter kind of normalisation may disappear. Therefore it seems to be a reasonable decision to stay with the current (first) normalisation technique which takes all F0 values over all meetings into consideration at once.

5.5.1.2 Naturally occurring speech

It is a requirement to use naturally occurring speech as the basis for interactional analyses, as it is used in the current study (cf. to Chapter 3). However, using natural speech can cause major problems for the acoustic analysis of interactional data (current chapter).

First, the utterances are not at all controlled – and not controllable, which means that it is almost impossible to compare the acoustics for example by restricting the analysis to identical words. In adjacent turn sequences of the type analysed in this study, one speaker hardly ever makes an utterance which is lexically identical to the utterance of the immediately preceding speaker. This means that fluctuations in the F0 contour (and possibly all other prosodic features), caused by the segmental production (intrinsic pitch), are different from turn to turn, from speaker to speaker. Because of the fact that the utterances are innately different, one would need to computationally eliminate the effect of the fluctuations. But this is already complicated under laboratory conditions and is expected to be even more complicated when it is tried to be implemented for naturally occurring speech.

Second, the speaking style is hardly anything that is similar to smoothly articulated and especially smoothly phonated turns. The amount of non-modal voicing is extremely high and causes problems, such as large portions of missing data. But, as there is no real equivalence to "missing prosody", it should be made possible to replace "missing" data by extracting other parameters which have not yet been explored in the current study, such as voice quality.

5.5.1.3 Missing data

Dealing with prosodic features, especially F0, means dealing with missing data. There are ways of working around this problem, such as interpolation, but the risk is to introduce errors that can't be accounted for at a later stage. Interpolation of missing F0 values has been successfully applied by Hermes (1998b) and Rilliard et al. (2011). But the data they have used is not taken from naturalistic interactional recordings. In the current study we avoided any attempt to interpolate, instead we modified the DTW algorithm. The final accumulative quality score was normalised by the amount of valid data in the utterance pair. This was shown to be successful when it was applied to artificial contours with gaps (see Section 5.3.2.4). However, if the amount of gaps increased above a certain level, the normalisation technique reached its limit.

5.5.2 False negatives and false positives

We have wished to find a clearer distinction between alignments and non-alignments with most of the similarity scores close to 1 for the alignments and most of the similarity scores close to 0 for the non-alignments. The distribution of the alignments gets relatively close to this expected pattern, but still shows many alignments with similarity scores close to 0. The distribution of the non-alignments is almost flat, i.e. has the same amount of similarity scores across the whole range from 0 to 1, except for the 0 itself, which is proportionally highly represented.

Table 11 collects the similarity scores that have been computed according to both metrics, the maximum similarity score (max_{Sim}) and the accumulative quality score (DTW_{Sim}). The table also shows what the prior and the target IPs of the adjacent turn examples are according to the Extracts in Chapter 4 (alignments on top and non-alignments at the bottom).

It can be seen that most of the turn pairs in the alignment category achieved relatively high similarity scores, which would mean true positives (correctly high similarity) and most of the turn pairs in the non-alignment category achieved relatively low similarity scores, which would mean true negatives (correctly low similarity). In this section we are however interested in the false positives (incorrectly high similarity) and false negatives (incorrectly low similarity) and try to explain why some non-alignments achieved unexpectedly high similarity scores and why some alignments achieved unexpectedly low similarity scores.

It is tempting to declare the cases which are deviant from the expected score as false negatives or false positives. But it is almost impossible to draw the line between positive and negative. Would all alignments with a score below 0.5 count as false negatives? Would all non-alignments with a score above 0.5 count as false positives? Or do we have to set the dividing score other than 0.5? The ROC-curve analysis used varying thresholds in order to estimate the sensitivity (true positive rate) and specificity (true negative rate) for a statistic classification attempt (see Section 5.4.3). This attempt is based on the recorded similarity scores and the interactional categories, which are treated as given. It shows that alignments and non-alignments can be statistically distinguished on the basis of these measures, but it does not explain *why* some alignments have unexpected low similarity scores and some non-alignments have unexpected high similarity scores.

5.5.2.1 False negatives

The most notable example for an alignment that unexpectedly achieved a low similarity score is the turn pair "`should be alright`" (prior IP) and "`uh-huh`" (target IP) from Extract 24 (see Table 11). Both similarity measures (max_{Sim} and DTW_{Sim}) achieved a score of zero. Looking at the prosodic contours of the two turns may help to explain this. Figure 42 illustrates the prosodic contours of the adjacent turns from the prior Speaker A (top) and the current Speaker B (bottom).

Table 11: Maximum similarity score (max_{Sim}) and normalised accumulative quality score (DTW_{Sim}) of the adjacent turn examples found in the extracts in Chapter 3. It is indicated what the prior IP and the target IP are. The examples of alignments (top) and non-alignments (bottom) are separated.

Extract	Prior IP	Target IP	max_{Sim}	DTW_{Sim}
		Alignments		
Extract 3	so	yeah	-	0.72
Extract 4	for that part	mm-hmm	-	0.05
Extract 5	purposes	yep	-	0.37
	purposes	yeah	-	0.51
Extract 7	really clear	okay	0.95	0.6
	exactly where the beep is	perfect	0.07	0.74
Extract 8	yeah	oh okay	-	0.6
Extract 9	just the parallelogram	yeah just the parallelogram	0.98	0.66
Extract 10	you know	it's not too bad	0.94	0.79
	bigger so is	uh-huh	0.21	0.47
Extract 11	but obviously they are like	they're the paid	0.7	0.64
Extract 12	the W3C	yep mm-hmm	0.93	0.83
	SVG	okay	-	0.83
Extract 13	slightly easier	mm-hmm	0.34	0.66
Extract 15	for doing it	uh-huh	0.25	0.37
Extract 16	to munich	oh yeah	0.92	0.49
	of december	mm-hmm	0.75	0.53
Extract 17	stuff	yeah	0.85	0.6
Extract 18	can't find anything	mm-hmm	0.24	0.15
Extract 20	JAST report	oh yeah	-	0.35
Extract 23	experiment	yeah	-	0.85
Extract 24	should be alright	uh-huh	0	0
Extract 26	so	right	0	0.35
	control panel	okay	-	0.12
Extract 30	so what i'm gonna	uh-huh	-	0.81
Extract 31	polygons now don't	oh that's good	0	0.62
		Non-alignments		
Extract 6	so	i mean	n.a.	n.a.
Extract 8	that was the eye	oh really	0	0.47
Extract 9	with rotation	just the parallelogram	-	0.45
Extract 10	bigger so is	i c*	0	0
Extract 12	files	so i wouldn't	0	0
Extract 13	slightly easier	uh yeah	0.01	0.52
Extract 14	the soundproof box	oops	-	0
Extract 16	on the friday	and john's going	0.55	0.42
Extract 20	on this project or something	what documentation	0.62	0.51
Extract 21	i think so	where are they	-	0.46
Extract 22	they'd given their	i bet marloes just talked to 'em	0.95	0.79
	cause i haven't had a reply back	ah i think	0	0.23
	just in case they had	is it hard	-	0.33
Extract 26	control panel	so we want one	0	0

In the final IP from Speaker A we can see a short fall from "be" to the beginning of "alright", followed by a rise from slightly below the speaker's mid range to slightly above it. The "uh-huh" from Speaker B similarly depicts a rising movement, however with a larger span and in a different range. The F0 spans over more than one standard deviation and is located relatively high in the speaker's range (1.5 standard deviations above the speaker's mid range). The low similarity scores can therefore be attributed to the high sensitivity of the algorithms to differences in the F0 range. This is evident from the low similarity values of the pairwise comparison of the F0 contours illustrated in Figure 43. The maximum similarity score (max_{Sim}) is very low and the accumulation of these low similarity values over the best alignment path (DTW_{Sim}) is equally low. Following the prediction that low similarity indicates non-alignment

leads here to a false negative as the interactional analysis identifies the "uh-huh" as an alignment.

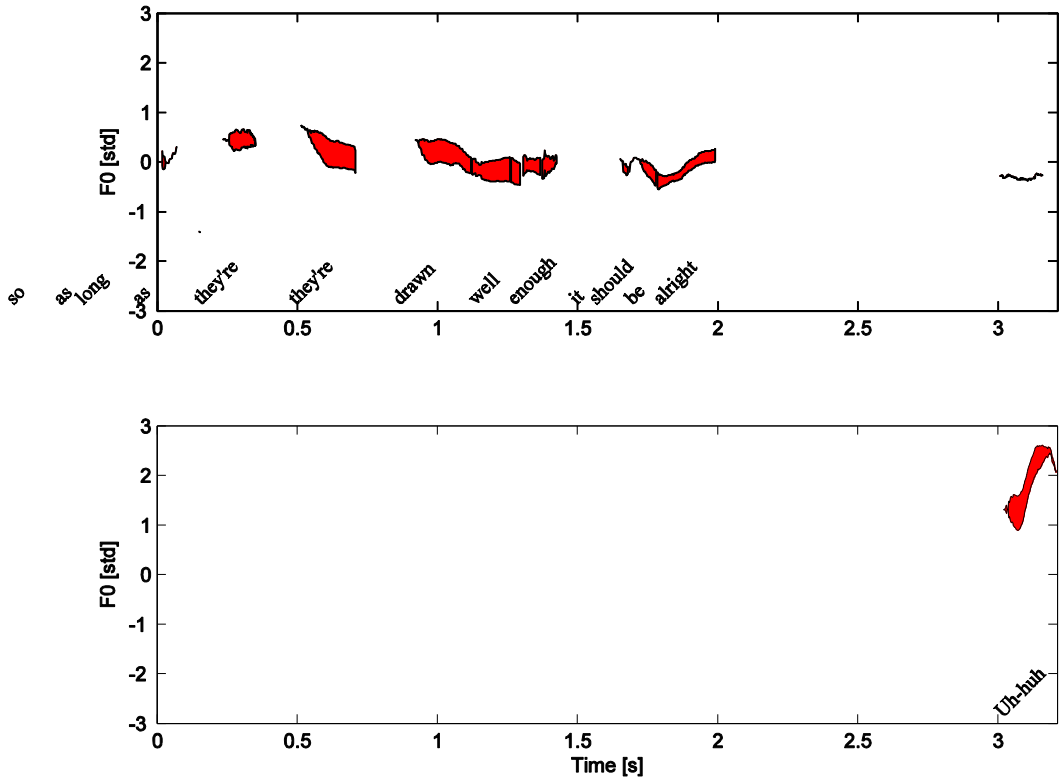


Figure 42: Intensity weighted F0 contours from Extract 24.

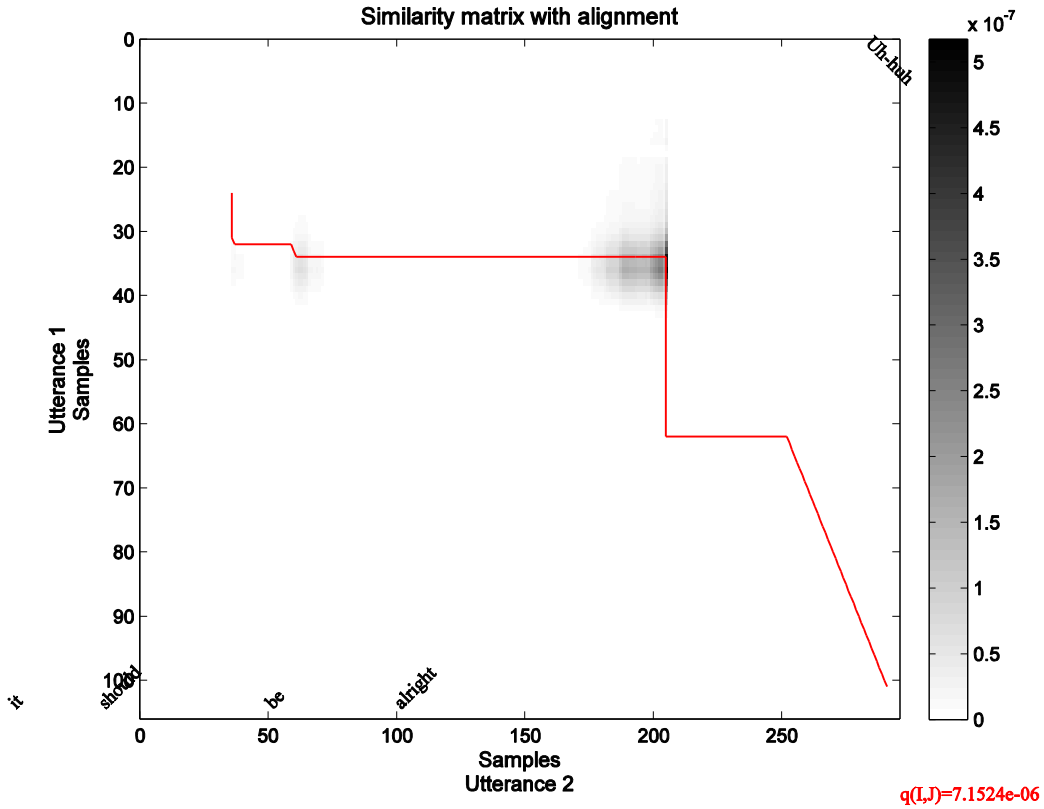


Figure 43: Similarity matrix of the comparison of the turn pair from Extract 24.

The answer to the question, why the prior speaker treated the turn from the other speaker as alignment with the need implied to continue on the own agenda, is speculative and manifold.

The relatively long stretch of silence (1 second) might have a special influence on the treatment of the following "uh-huh". Or the fact that the second speaker nods simultaneous to the verbal "uh-huh" and is resumed even after might be the decisive cue. It is also possible that the match of the rising F0 movement was sufficient and the mismatch in the F0 range did not matter.

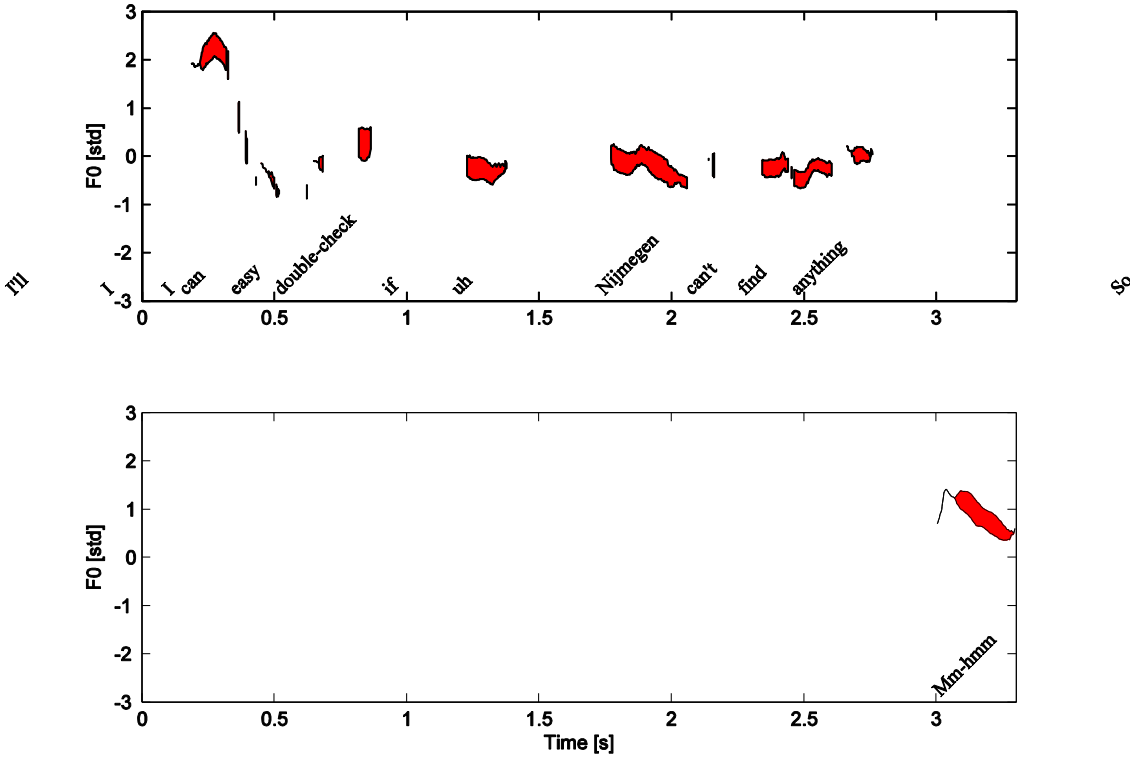


Figure 44: Intensity weighted F0 contours from Extract 18.

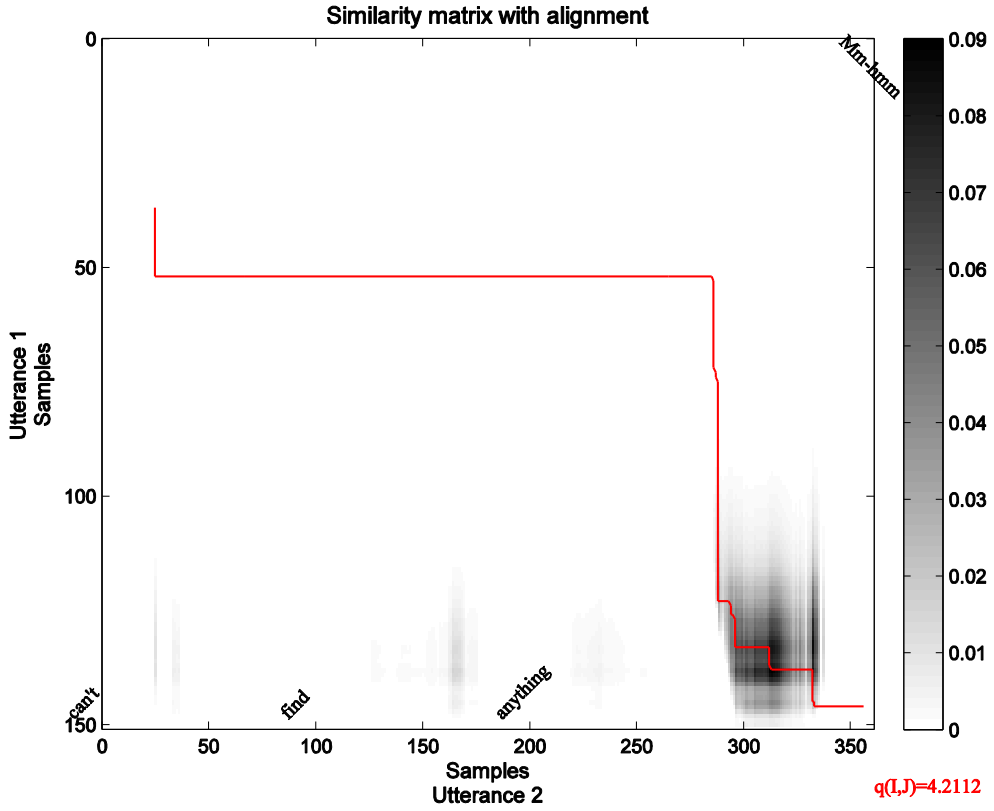


Figure 45: Similarity matrix of the comparison of the turn pair from Extract 18.

A second alignment example with a relatively low similarity score is the turn pair "can't find anything" (prior IP) and "mm-hmm" (target IP) from Extract 18 (see Table 11). The maximum similarity score is 0.24 and the accumulative quality score is 0.15. Figure 44 illustrates the prosodic contours of the prior turn from Speaker A (top) and the following "mm-hmm" from Speaker B (bottom). While Speaker A's F0 is around his mid range and the F0 movement on the last three syllables ("anything") is progressively rising, Speaker B's F0 is approximately one standard deviation above her mid range and constantly falling. These differences have also influenced the similarity matrix which has been computed in the comparison of the two turns, illustrated in Figure 45.

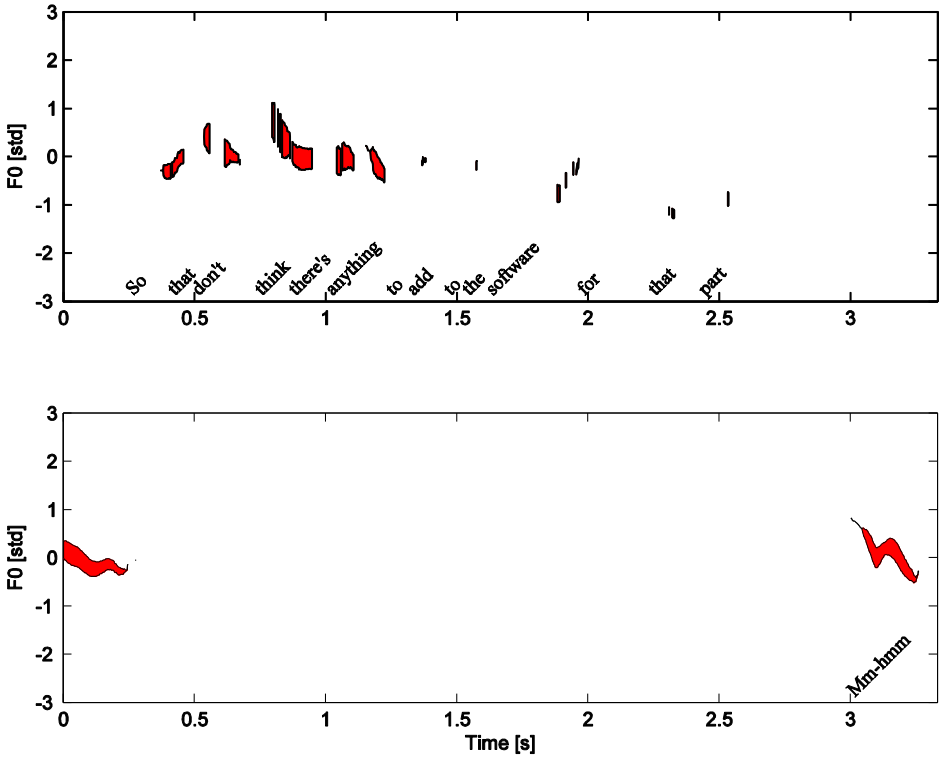


Figure 46: Intensity weighted F0 contours from Extract 4.

Although these differences are not as big as in the previous example, the similarity values are relatively low. Which of the two, the difference in the F0 movement, or the difference in the F0 range had a bigger influence on the similarity score in this case is not sure in the end. But the difference in the similarity scores between max_{Sim} (0.24) and DTW_{Sim} (0.15) can give rise to the speculation that the F0 movement difference is better modelled by the latter, as it accumulates the similarity values in the similarity matrix over time. So, as the F0 movement is nothing else than a change over time, differences in that movement will be penalised and lead to a lower overall score than the pure search for one maximum. In the current example this has, however, led to a lower similarity score and worsen the case for the alignment.

A third and last example of an alignment with a low similarity score that should be discussed here is the turn pair "for that part" (prior IP) and "mm-hmm" (target IP) from Extract 4 (see Table 11). While no similarity score could be computed for max_{Sim}, DTW_{Sim} is relatively low with a value of 0.05.

The prosodic contours of the prior turn from Speaker C (top) and the following "mm-hmm" from Speaker B (bottom) are illustrated in Figure 46. Here, the contour of the second turn has the shape of a fall-rise-fall in the mid of the speaker's range. But in the contour of the prior speaker's last IP ("for that part"), hardly any F0 readings are available. These are not enough for the computation of max_{Sim} and the few values which are available for the estimation of DTW_{Sim} (see Figure 47) make the measure of a score of 0.05 unreliable.

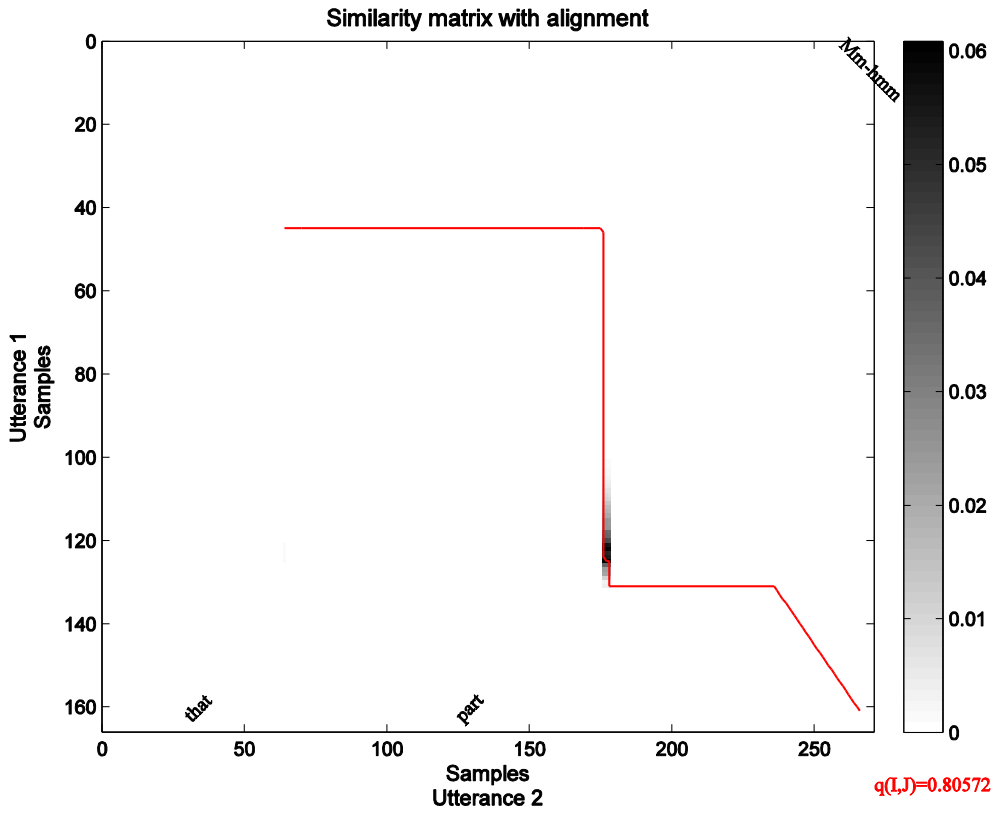


Figure 47: Similarity matrix of the comparison of the turn pair from Extract 4.

5.5.2.2 False positives

An example of a non-alignment which scored unexpectedly high similarity scores in max_{Sim} (0.95) and DTW_{Sim} (0.79) is the turn pair "they'd given their" (prior IP by Speaker C) and "I bet Marloes just talked to 'em" (target IP by Speaker B) from Extract 22 (see Table 11). Figure 48 illustrates the prosodic contours of the two turns. The contour of the prior turn starts in the mid range for "they'd" and falls slightly below it for the rest ("given their"). A similar pattern can be observed for the part "Marloes just talked to 'em" for the second speaker. But the contour on the first bit of it ("I bet") resembles a step up from "I" to "bet" and is located above the speaker's mid range. In the similarity matrix (Figure 49), the best alignment path skips the first bit of it and aligns "Marloes just talked" with "they'd given" and "to 'em" with "their". Because those parts fit quite well, the overall similarity score is relatively high. However, the accumulative quality score (0.79) is not as high as the maximum similarity score (0.95).

The reason why the prior speaker chose to treat the turn by the second speaker as an initiation to which it is necessary to orient to in a specific way, is again speculative, but there are indications that show special treatment of the target turn in the successive turn. The second speaker's turn is relatively long and contains a syntactically complete sentence with a strong formulation of the speaker's stance ("i bet ..."). It introduces a comment on the report of the prior speaker on what he was doing. He wrote an email to Nijmegen, a joint location of the project, to ask if a "set of instructions" for participants was used at some stage of the conducted research. But if someone (Marloes, who works in Nijmegen and who is the suspected addressee of the email by C) "just talked" to the participants, such an email seems useless in retrospect. Therefore, the prior speaker might have chosen to orient in his successive talk to this comment with a nod and a truncated "it probably i s*". He finishes with "i suspect so because i haven't had a reply back", indicating orientation to the target IP that needs special treatment.

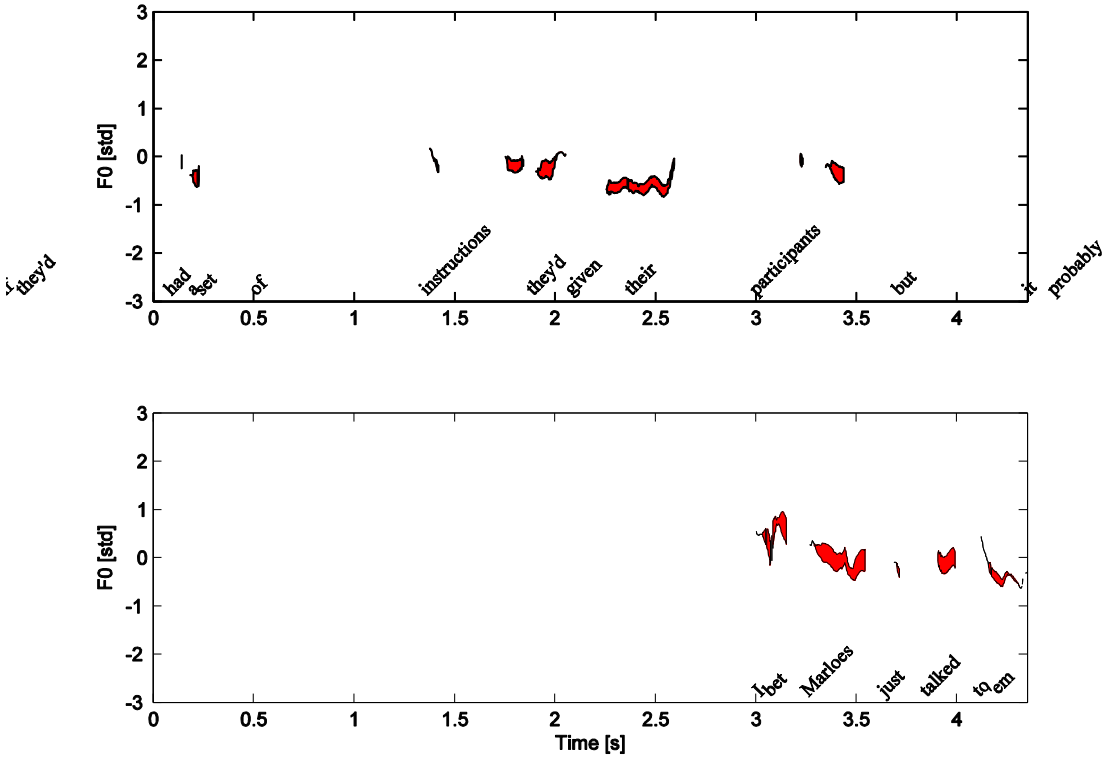


Figure 48: Intensity weighted F0 contours from Extract 22.

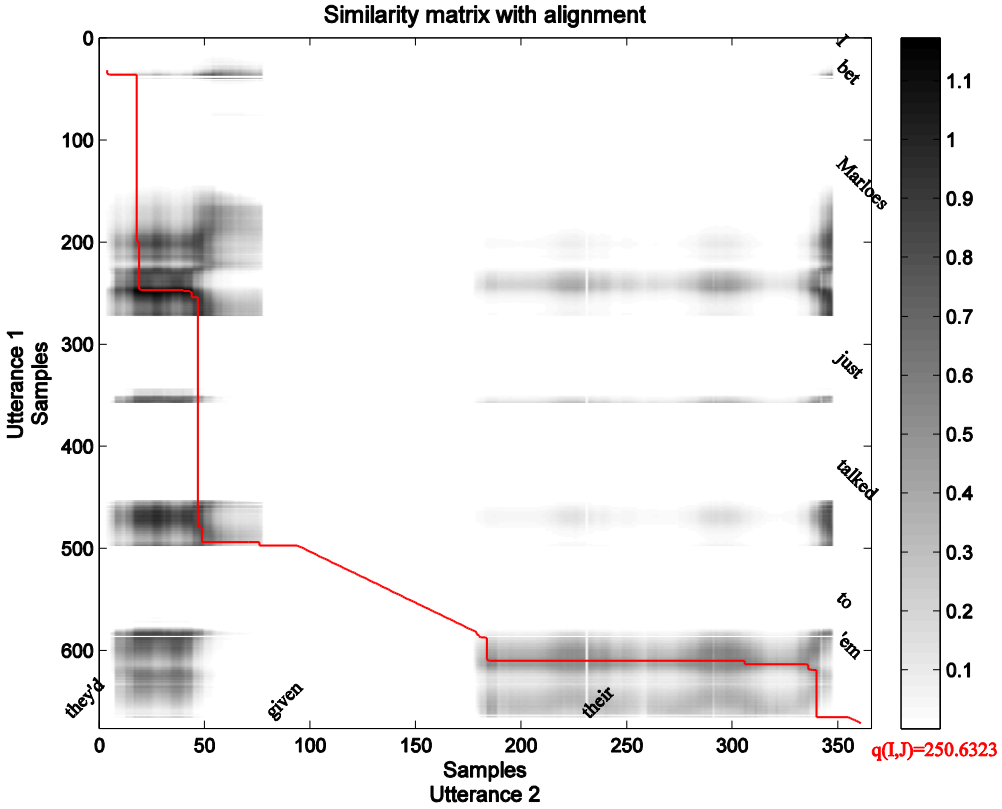


Figure 49: Similarity matrix of the comparison of the turn pair from Extract 22.

The lexico-syntactic features seem to outweigh in this case the prosodic matching features. However, the first bit of the second speaker's turn does not match the prior speaker's turn that much. Therefore, the high similarity score might just be an artefact of the segmentation into intonational phrases. The non-match of "`I bet`" might already do the non-aligning work and

would need to be separated from the reminder of that turn. The overlapping region with the prior speaker might also have some influence on the organisation of the interactional sequence, but we leave it for future research.

A second example of a non-aligning turn which achieves relatively high similarity scores (max_{Sim}: 0.62 and DTW_{Sim}: 0.51) comes from Extract 20 (see Table 11). The turn "`on this project of something`" (prior IP) from Speaker C is followed by "`what documentation`" (target IP) from Speaker B. The prosodic contours are illustrated in Figure 50. Although the prior speaker (top panel) speaks in a creaky voice quality, some of the F0 values are just below the speaker's mid range. Those F0 values in the last words ("`or something`") are above the speaker's mid range and produced with very low intensity. The second speaker's first word "`what`" is very high, with three standard deviations and more above the speaker's mid range. For the first syllable of the second word "`documentation`" the F0 falls back to mid range and for the following syllables even below. These overall prosodic patterns seem to contrast each other, which is also depicted in the similarity matrix in Figure 51.

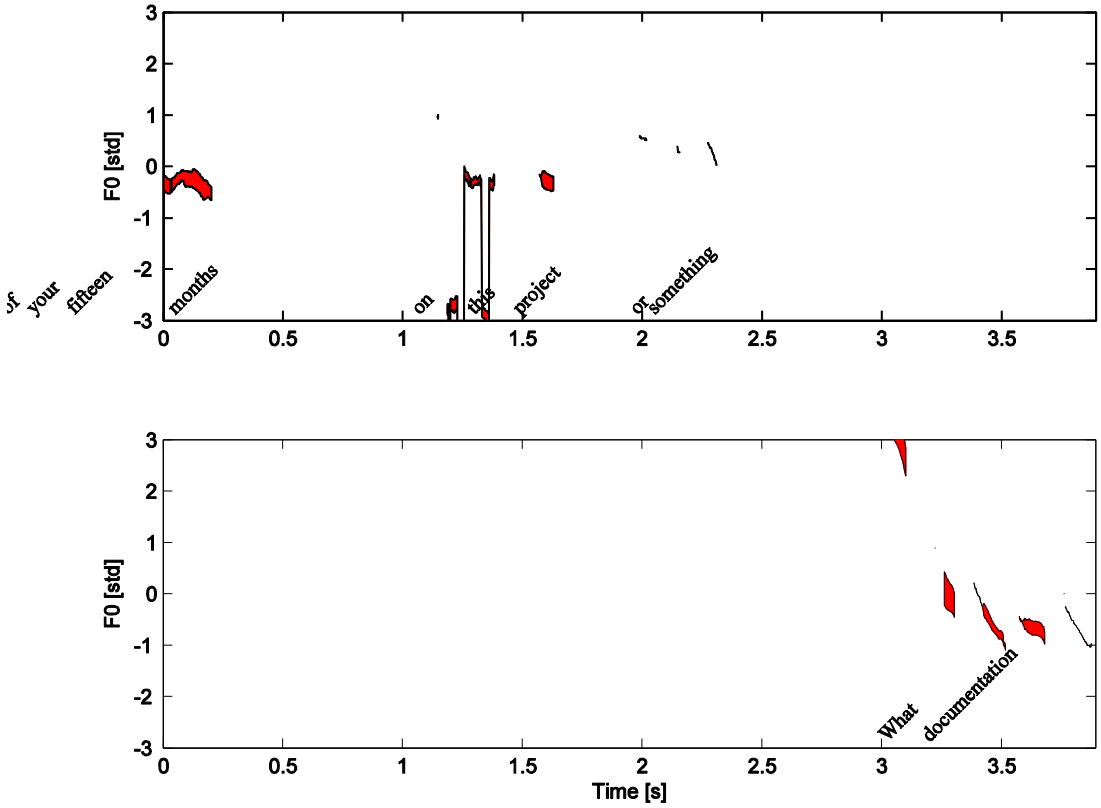


Figure 50: Intensity weighted F0 contours from Extract 20. The almost vertical lines in the top panel between second 1 and 1.5 are artefacts from the algorithm which creates the visual representation of the intensity weighted F0 contours and struggles here with the creaky voice quality. In the bottom panel, the high F0 on "what" exceeds three standard deviations of the speaker's mid range.

The "`what`" from the second speaker does not match the prior talk and is skipped by the DTW algorithm (vertical alignment path in the top left corner). The same happens with the "`or something`" from the prior speaker (horizontal alignment path in the bottom right corner). But the beginning of the prior talk is matched with the end of the second turn quite well. It is due to those parts that the overall similarity scores are relatively high. Similar to the "`i bet`" in the previous example, it can be argued that the extreme non-match of the prosody of "`what`" does already most of the non-aligning work, while the rest of the turn is released from that job. It even gives rise to think about the segmentation of the underlying units into IPs.

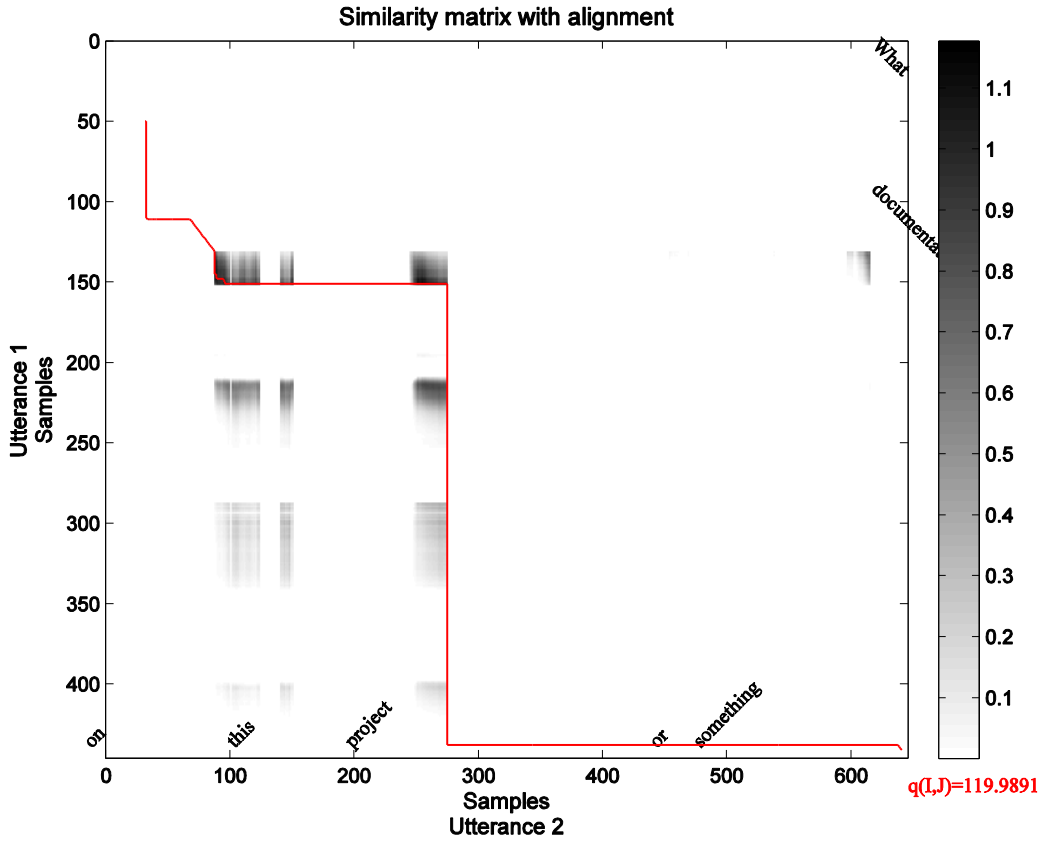


Figure 51: Similarity matrix of the comparison of the turn pair from Extract 20.

The last two examples, in which some parts of the contours match each other and some parts don't, indicate that some finer grained unit than the intonational phrase would be required in the analysis of alignments and non-alignments with respect to prosodic matching.

5.5.2.3 Differences between max_{Sim} and DTW_{Sim}

Sometimes the difference in the measures of the two similarity metrics is rather big – not only if one of the scores could not be computed at all, while the other provided a number for the similarity – but also if both provide a number. For example in the turn pair "`where the beep is`" (prior IP) and "`perfect`" (target IP) from Extract 7 (see Table 11), the two measures diverge strongly with max_{Sim} of 0.07 and DTW_{Sim} of 0.74. Another example is the turn pair "`slightly easier`" (prior IP) and "`uh yeah`" (target IP) from Extract 13 with max_{Sim} of 0.01 and DTW_{Sim} of 0.52.

The prosodic contours of the latter example are illustrated in Figure 52. The contour of the prior Speaker A (top panel) is located approximately one semitone below the speaker's mid range for the last voiced sound of "`slightly`" and the first voiced sound of "`easier`" and rises to the end up to the mid range. The contour of the second Speaker C (bottom panel) is located around the mid range and slightly falls during the "`yeah`" ("`yes`" in the figure). From these patterns it can be expected that the contours match in F0 height quite well at the end of the prior turn and less good at the beginning (disregarding the beginning of the word of the "`slightly`", as no F0 readings are available for that stretch).

The relatively poor match in F0 height at the beginning of "`easier`" is also reflected in the similarity matrix which was computed with the maximum similarity search algorithm (see Figure 53). Those regions at the end of the word "`easier`" which would match better did not have enough F0 readings and their similarity could not be computed. Therefore the search algorithm found only the low maximum similarity (0.01) in the whole matrix.

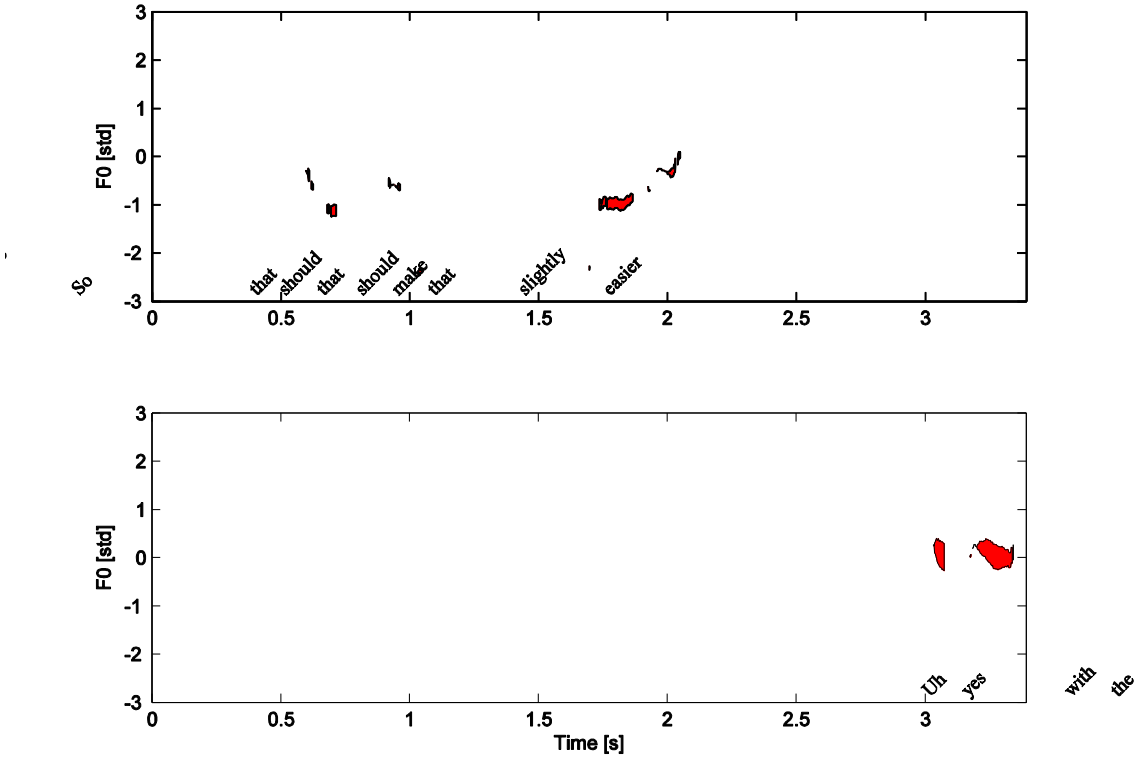


Figure 52: Intensity weighted F0 contours from Extract 13.

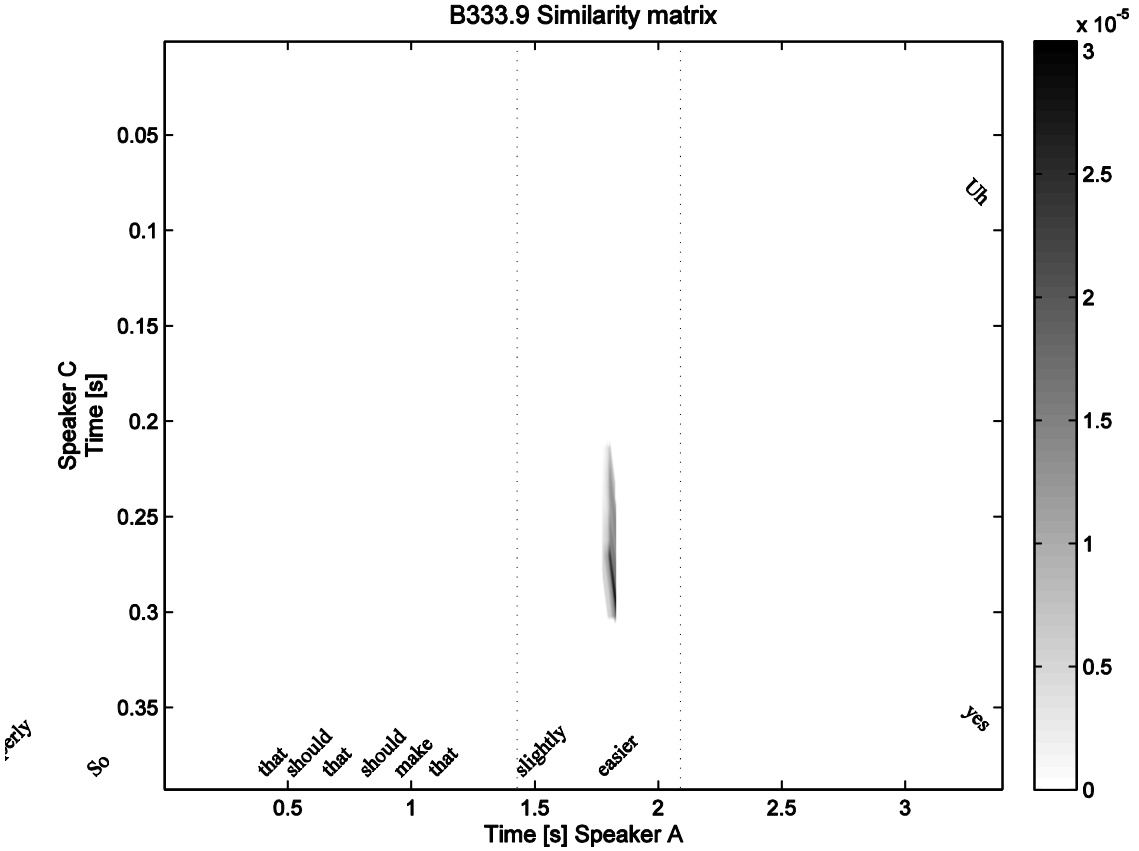


Figure 53: Similarity matrix of the comparison of the turn pair from Extract 13 using the maximum similarity search algorithm.

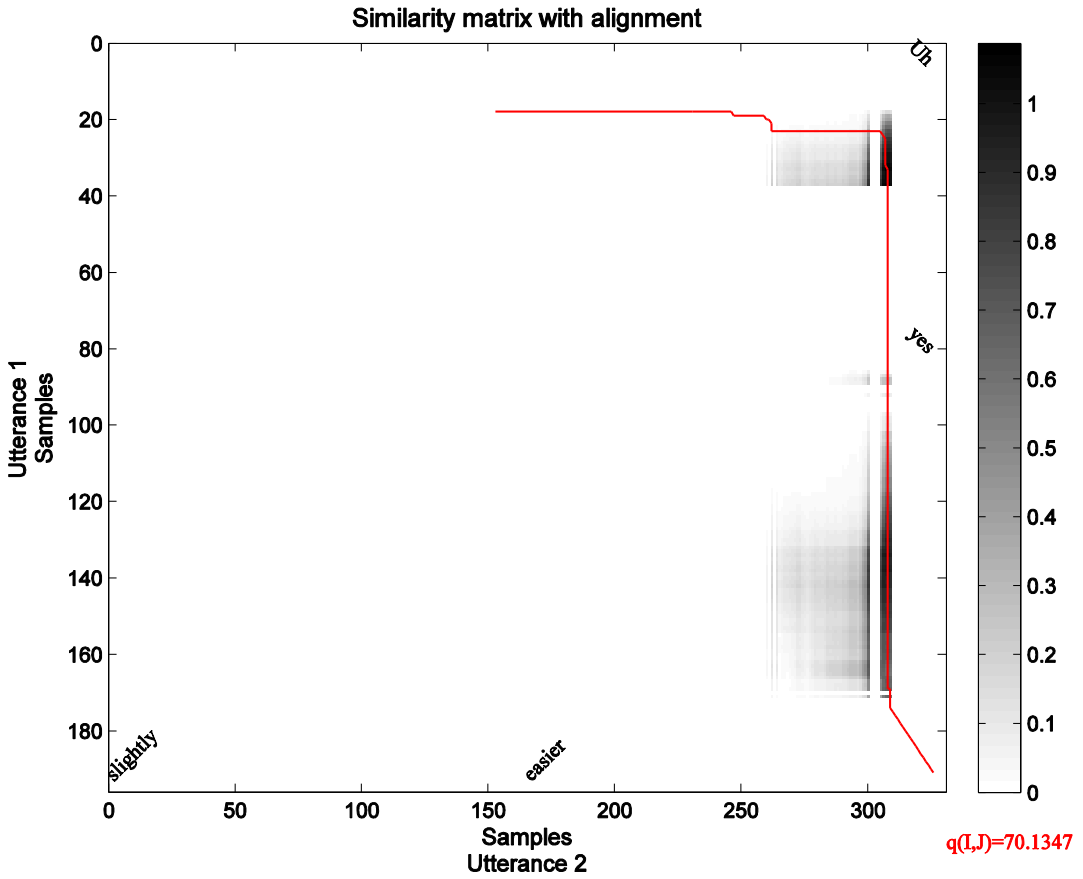


Figure 54: Similarity matrix of the comparison of the turn pair from Extract 13 using the accumulative quality score.

The sparseness of F0 values is less a problem for the DTW algorithm (see Figure 54). It computes the similarity of each available F0 value pair. The best match is found for the second speaker's turn "uh yeah" with the very end of the prior turn. The best alignment path accumulates these similarity scores and reaches a relatively high overall similarity score of 0.52.

It is arguable, which of the two measures now better reflects the acoustic reality. But apart from the acoustics, the interactional reality also needs to be addressed. If the prosodic matching hypothesis holds, we would expect a low prosodic similarity score for the current example, as it has been classified as non-alignment on interactional grounds. On the other hand, the relatively high similarity score measured with the DTW algorithm may give rise to re-think the interactional category. Why was the "uh yeah" classified as non-alignment? The reasons were that the prior speaker did not continue speaking or gesturing along the lines of the prior agenda, no pause developed after that turn and the current speaker immediately latches into the next IP ("with the um") without leaving space for the prior speaker to resume. This classification could be called into question, as an alternative categorisation could be to classify the "uh yeah" as alignment and the subsequent "with the um" as non-alignment which initiates the new agenda.

In summary, there are several possible explanations why we observe the unexpected distributions of similarity scores which appear here.

- It is possible that the algorithm that computes the similarity between the contours is not optimal.
- The implementation of the prosodic parameters (here F0 only) might not be sufficient to model all the details that interactional participants can draw their attention to.
- Gesture might be one further such parameter, additional to other prosodic parameters (loudness, voice quality, etc.).

- The segmentation of the speaker's turns into units of intonational phrases might not have been the best choice. A finer grained structure of smaller chunks might improve the analysis.
- It is possible that the interactional classification would require more than the two suggested categories to indicate the participants' actions, for example different levels of cooperativity.
- The interactional classification of some examples might be faulty and the annotator might have missed one or some relevant interactional cues.
- The prosodic acoustic properties may not be employed in a discrete way for matching or non-matching the other speaker, but may be employed more gradually on a scale from identical match to extreme non-match.
- It is difficult to distinguish honest utterances from sarcastic utterances if no subsequent laughter or other indications are present. It is possible that a second turn matches prosodically with the prior turn, which would indicate alignment. But if the speaker uses this technique ironically, such a match can also cause a subsequent special treatment by the prior speaker, e.g. laughter.

6 Gestural analysis and prosodic-gestural model

This chapter addresses the gestural properties of conversations. It is split into two parts. In the first part, the research questions on the gestural properties are addressed. It is investigated whether the individual gesture of the target speaker is responsible for the treatment of the target turn as alignment or non-alignment or whether gestural matching is the driving factor for the status of the social action. In the second part, gestural properties are compared with prosodic similarity in a prosodic-gestural model. It is investigated whether the two modalities of gesture and prosody interact to determine the social actions of alignment and non-alignment.

6.1 Gestural analysis

Research on gesture and speech suggests that the two modalities are tightly linked (Kita, et al., 2007). Gesture facilitates the production process of speaking (Habets, Kita, Shao, Özyürek, & Hagoort, 2011) and is also helpful on the recipient's side in the process of comprehension of spoken utterances (Holle, et al., 2012). These studies focus mainly on the synchronisation of speech and manual gestures. Other studies take head movements, body movements and facial displays into account. Such gestures are also considered to have a strong influence on discourse and the management of social actions (Ekman & Friesen, 1969). Some of these gestures are used by conversational participants as imitating gestures (Scheflen, 1968; Wundt, 1973). Interactional studies have shown that gestures are used to perform specific social actions (Schegloff, 1987; Stivers, 2008; Whitehead, 2011). The gestural context is considered to have a strong influence on other gestures and it is considered that this gestural matching or non-matching is used as a resource to perform specific social actions (Lerner, 2002; Selting, 2010).

Several questions and hypotheses arise concerning gesture in relation to the social actions of alignment and non-alignment that have also been the focus of the previous chapters. The questions addressed below ask whether alignments and non-alignments are accompanied by specific gestures or alternatively, by gestural matching and non-matching.

The first research question asks if alignments and non-alignments are performed with specific gestures (RQ3a), i.e. whether the gesture in the second turn can *itself* help determine the interactional category. The second research question asks if alignments are performed with gestural matches and non-alignments with gestural non-matches (RQ3b), i.e. whether of the immediately preceding gesture of the prior speaker.

In the interactional analysis (Chapter 4), the sequential organisation of gestures has already played a role in determining the treatment of a turn at the second position as alignment or non-alignment. Thereby, the fact that the prior speaker's gesture was continued successively, was used as cue. This was interpreted as a continuation of the prior speaker's (gestural) agenda, which indicates that the interactional category of the target turn was an alignment. Non-aligning target turns rarely make the prior speaker continue on the prior gestural agenda.

In this chapter, we intend to analyse whether the gesture in the second turn *itself* can help determine the interactional category, or if the interactional category depends mainly on gestural matching or non-matching of the immediately preceding gesture of the prior speaker.

It should be noted that there is a risk of some circularity in the procedure, as all gestural layers were available to the annotators through the orthographic transcripts at the stage of interactional categorisation. This may have influenced the decisions of the annotators to some degree. However, the classification instructions explicitly ask to use the gestures of the prior speaker.

The gestures of the target speaker were not mentioned at all in the classification instructions. Although the second annotator may attest that the gesture of the target speaker did not influence her decision, an influence may still have taken place unconsciously.

There would have been an alternative to this procedure which is to omit the gestural transcription in the same way as the prosodic transcription was omitted for reasons described in Section 3.1.5. However, if the gestural transcripts were omitted, the annotators would not have had the chance to refer to the gestures used by the prior speaker in the prior turn and the successive turn. This was one of the classification criteria, which can be regarded as being outside this risk of circularity.

Nevertheless, the focus of this thesis is on the approach taken for the analysis of gestures in their sequential and in their contextual use, namely in relation to social actions. Even if annotators may have been influenced unconsciously by gestures employed by the target speaker during the target turn, a post analysis (below) of those gestural properties can still provide useful information about the organisation of the social actions by gesture, as no relationship between specific gestures and specific social actions has been suggested beforehand.

6.1.1 Research questions and hypotheses

The review of literature on gestures in adjacent turns reveals two fundamentally different concepts. First, it is possible that the second turn gains its communicative function from the co-speech gesture itself. This means that a certain gesture in co-use with certain words functions in a specific way. For example, a certain gesture might mark alignment, allowing the prior speaker continue on the prior agenda, alternatively, it might function as a non-alignment, making special treatment of that turn by the prior speaker relevant. Second, it is possible that the second turn gains its communicative function from the co-speech gesture, depending on its relative match with the prior speaker's gesture.

With respect to the current collection of alignments and non-alignments and with respect to the literature discussed above and in the review Section 2.2.5, this leads to the following two specific questions according to the above stated general research questions (RQ3a and RQ3b):

RQ3a: Are alignments and non-alignments performed with specific gestures?

In other words: Does the interactional category of the target turn depend on the accompanied gesture itself?

RQ3b: Are alignments performed with gestural matches and non-alignments with gestural non-matches?

In other words: Does the interactional category of the target turn depend on a gestural match of the two adjacent gestures?

Following the assumptions prevalent in the literature, i.e. that gestures can perform specific social actions, one would expect to find each interactional category preferably associated with a specific gesture, or group thereof. For example, one could hypothesise that alignments are mostly performed using nods rather than any other gesture. The null hypothesis would imply that there are no specific gestures accompanying turns of one or the other category. The research question that is addressed here is RQ3a.

Assuming that gestures perform social actions according to their match or mismatch with the prior speaker's gesture (in the prior turn), one would expect a higher degree of matching between prior and target gestures in one category than in the other category. For example, one could hypothesise that alignments are indicated by a gestural match, rather than a mismatch.

The null hypothesis would then represent the notion that the interactional categories are independent from the degree of match of the two gestures. In this case, there will be low matching between prior and target gestures. The research question that is addressed here is RQ3b.

6.1.2 Methodology

The gestures of the prior turn and the target turn were recorded as described in Section 3.1.3. In order to address the research questions of this first part of Chapter 6, two approaches are compared. The corresponding hypotheses are tested statistically. Here, the gestural domain was analysed independently of the auditory domain. The first hypothesis claims that the interactional category of the target IP only depends on the gesture of the same speaker that accompanies the target IP. The second approach claims that the interactional category of the target IP depends on the match of the target speaker's gesture with the preceding gesture of the prior speaker.

The records of gestures contain only the main gesture types, neglecting the detailed shapes and movements of the according body parts. It is therefore neither possible to determine the real gesture (e.g. the exact shape of the hand during hand gesticulation), nor whether it was an exact match of the prior gesture. Gestures and gestural matches can merely be approximated through their overall gesture types. Errors might arise from this limitation, but it was decided to accept these in order to generate a larger amount of data, which can be used for statistical analyses. (A device for tracking exact body movements precisely could help to determine gestural matches. A prototype is described in the last chapter of this thesis.)

Only data from meetings B and C were used. This was due to the recording set-up in meeting D, which made the data from this meeting unsuitable for analysis. For example, participants frequently moved to the white-board, leaving the viewing angle of the camera. Gestures were no longer recorded. The same applies to those participants who remained seated, but turned their back to the camera, which makes gesture annotation impossible.

According to the research questions and hypotheses formulated above, two types of tests are necessary:

The first test requires an evaluation of whether a single gesture or multiple gestures are used significantly more often for alignments than for non-alignments. It was anticipated that nods were more likely to contribute to alignments than to non-alignments.

The chi-square test was chosen for statistical analysis because the underlying data is recorded on a nominal scale and therefore represents frequencies. The testing of the first hypothesis, which says that the current gesture alone determines the interactional category, is described in Section 6.1.3.

The second test requires an evaluation of the matching of all pairs of gestures, i.e. whether the prior gesture (from the prior speaker) and the current gesture (from the target speaker) are the same. It can then be established whether the rate of matching is significantly different for alignments and for non-alignments. Matching would be expected to be higher for the alignment group than for the non-alignment group.

The appropriate statistical test for the matching of adjacent gestures is a three way chi-square test. Cohen's Kappa is useful for the description of the data; Kappa was found to be adequate for this kind of analysis (cf. to agreement measures in Section 4.3.2). The testing of the second hypothesis, which states that the interactional category is determined by the match of the current gesture with the gesture of the prior speaker, is described in Section 6.1.4.

6.1.3 Determining interactional category from gesture type

This section addresses research question RQ3a. The hypothesis is that specific gestures determine the interactional categories alignment and non-alignment. The data used for the gestural analysis is the same as for the acoustic analysis. At each first and second turn pair, the predominant gesture which the according speaker performed was recorded. Therefore the number of gestures of the target speaker is equivalent to the number of oral turns in the two meetings under investigation. The overall number of target turns and therefore the overall number of target gestures is 812, where 162 come from speaker A, 436 come from speaker B and 214 come from speaker C.

6.1.3.1 Results

Table 12 shows the distribution of gesture types for each individual speaker. The same data is presented graphically in Figure 55. These are all gestures that coincide with speech. There are no turns without spoken material, as the criterion for the collection of instances was that every target turn (speech) needs to be preceded by some talk from another talker (cf. to the list at the end of Section 3.1.4). After the target turn, stretches of silence with potential productions of gestures may occur that can count as separate turns (cf. Schegloff (1987) and see Section 2.2.5.2). But this does not apply here, as only the prior turn and the target turn are at issue.

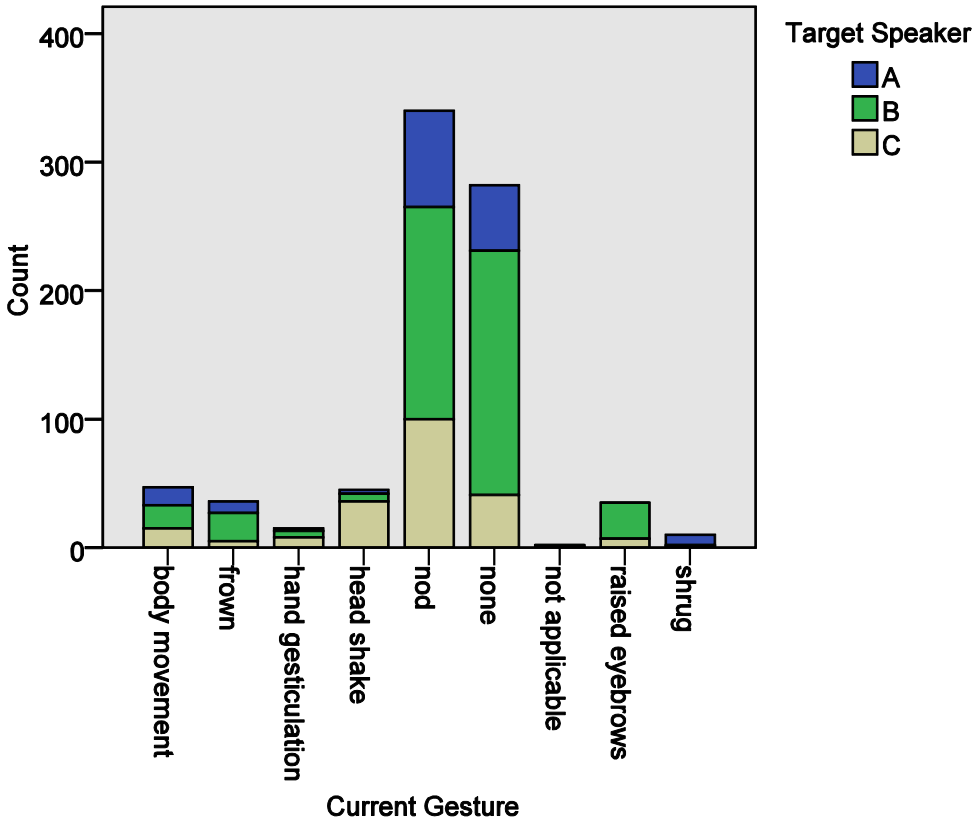


Figure 55: Distribution of gesture types in the target IP (current gesture) according to speakers. Nods are most frequent. Second most frequent is the category "none", followed by body movement, head shake, frown, raised eyebrows, hand gesticulation and shrug. In two cases the video was faulty and therefore no gesture type could be allocated ("not applicable"). Most of the raised eyebrows originated from speaker B and most head shakes from speaker C.

Overall, the most frequent gesture type is the nod (41.9%) followed by the "none" category (34.7%). These two make up 76.6% of all gestures types. The remaining gestures are shared between the other categories, which are body movement, hand gesticulation, head shake, frown, raised eyebrows and shrug. In two cases, no gesture type could be allocated due to faults in the video stream. These were included in Table 12 using the "not applicable" tag.

There seems to be a speaker dependency on what gesture is most often performed. For example, while speaker A never used raised eyebrows in target speaker position, most raised eyebrows came from speaker B (28) and a few from speaker C (7). On the contrary, body movements seem to be rather equally distributed between the three speakers with 14, 18 and 15 occurrences for speaker A, B and C respectively. But this relates to only 4.1% of speaker B's gestures. It would be interesting to find a method to take into account the normal behaviour of the individual, as some people tend to nod more or use more raised eyebrows. However, this is beyond the scope of this thesis.

Table 13 summarises the gesture types of the target IP for the social actions alignment and non-alignment. A graphical representation is given in Figure 56.

Table 12: Absolute and relative numbers of gestures of the target speaker during the target IP.

Current Gesture		Target Speaker							
		A		B		C		A, B and C	
		Count	%	Count	%	Count	%	Count	%
body movement		14	8.6	18	4.1	15	7.0	47	5.8
frown		9	5.6	22	5.1	5	2.3	36	4.4
hand gesticulation		2	1.2	5	1.2	8	3.7	15	1.8
head shake		3	1.9	6	1.4	36	16.8	45	5.5
nod		75	46.3	165	37.8	100	46.7	340	41.9
none		51	31.5	190	43.6	41	19.2	282	34.7
not applicable		0	.0	1	.2	1	.5	2	.2
raised eyebrows		0	.0	28	6.4	7	3.3	35	4.3
shrug		8	4.9	1	.2	1	.5	10	1.2
All	Count	162		436		214		812	
	%		20.0		53.7		26.4		100.0

Table 13: Current gesture according to the interactional category of the target IP.

Current Gesture	Interactional Category				Total	
	alignment		non-alignment			
	Count	Percent	Count	Percent	Count	Percent
body movement	16	3.4%	31	9.3%	47	5.8%
frown	7	1.5%	29	8.7%	36	4.4%
hand gesticulation	5	1.1%	10	3.0%	15	1.9%
head shake	19	4.0%	26	7.8%	45	5.6%
nod	285	59.9%	55	16.5%	340	41.9%
none	131	27.5%	151	45.2%	282	34.7%
not applicable	1	.2%	1	.3%	2	.3%
raised eyebrows	8	1.7%	27	8.1%	35	4.3%
shrug	5	1.1%	5	1.5%	10	1.2%
Total	476	100.0%	334	100.0%	812	100.0%

Some gesture types are more frequent in one category than in the other. Most prominent are nods in the alignment category. Nods occur approximately twice as often as the second category, "none". Some head shakes and body movements occur, but not many more than raised eyebrows, frowns, shrugs and hand gesticulations, which are sparse. Shrugs and hand gesticulations are equally sparse in the non-alignment category. More frequent are body movements, frowns, raised eyebrows and head shakes. Although the overall number of non-alignments is smaller than that of alignments, these specific gestures are notably more frequent in the non-aligning category. Nods still outnumber the just mentioned gesture types, even in the non-alignment category, but they still occur when compared with the alignment category. Most likely in the non-alignment category are cases of no gesture at all ("none").

It could be argued that some gesture types in Table 13 might be grouped together. For example frown and head shake are broadly speaking "negative". Given the way preference organisation works, i.e. one action is a preferred second pair part of a specific first pair part (Schegloff, 2007), one could expect certain asymmetries between speakers' production of different gesture types. However, as described earlier (Section 3.1.3), it was decided to avoid any such assumptions that associate specific forms, e.g. frown, head shake or specific prosody, with specific functions, e.g. the negative. Note also that several actions might be performed with head shakes: to express the negative, but also disagreement and intensification (Schegloff, 1987). Other groupings would be even more questionable, such as body movement with shrug

or raised eyebrows with hand gesticulation. Therefore it was decided to keep the gesture types separate.

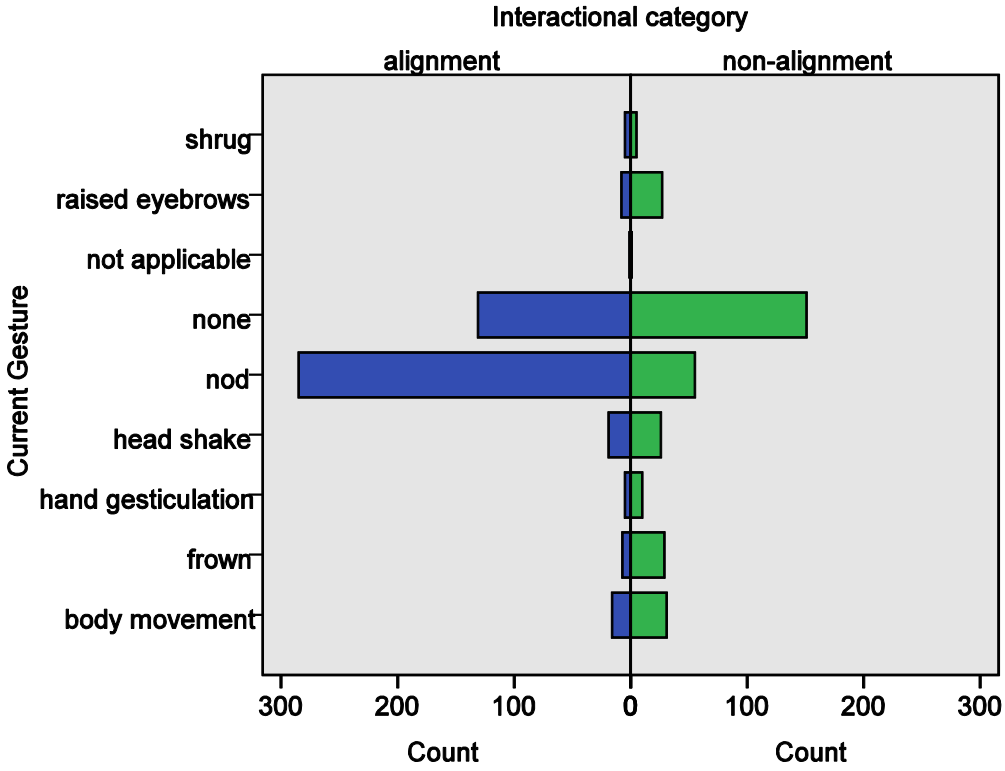


Figure 56: Gesture during the target IP according to the interactional category. In the alignment category, nods appear most frequently. Second most frequent is the category "none", followed by body movement, head shake, frown, raised eyebrows, hand gesticulation and shrug. In the non-alignment category, the largest group in absolute numbers were the "none gestures", followed by nods. The following gestures occur with approximately the same frequency: body movement, frown, head shake and raised eyebrows. Less frequent are shrug and hand gesticulation.

The first hypothesis claims that certain gestures are used by the target speaker during the target turn to align or non-align with the prior speaker. The distribution of current gestures as shown in Figure 56 would suggest that the interactional category is determined by the current gesture. The gesture nod seems to accompany alignments disproportionally more often (59.9%) than non-alignments (16.5%). The categories body movement, frown, head shake, hand gesticulation and raised eyebrows accompany non-alignments more often than alignments. When no gesture accompanied the target IP, this occurred as frequently in cases of alignment as in cases of non-alignment. However, the relative numbers indicate that non-alignments are accompanied more often (45.2%) by the "none" gesture category than alignments (27.5%).

6.1.3.2 Statistical test and discussion

A chi-square test was used to determine whether the interactional categories alignment and non-alignment correlate with the current gesture. According to Nachtigall and Wirtz (2004, p. 172), the chi-square test requires that the expected frequency of each gesture combination (see Table 13) exceeds 5. This is the case for all cells in the crosstabulation, except for the combination of shrug and non-alignment, where the expected frequency was 4.1. Because this is only a small deviation from the required frequency in one of 16 cells, it was decided that the data was suitable for processing with the chi-square test.

There was a significant interaction between the alignment category and the current gesture type: $\chi^2(7, N = 810) = 1.686E2$, $p < 0.001$. This result shows that alignments and non-alignments were produced using a different set of gestures. The effect size was measured using Pearson's phi ($\phi = \sqrt{\chi^2/N}$). It takes the square root of the chi-value from the chi-square test divided by N, the total sample size. Here it was 0.46.This indicates that the gesture type used by the target speaker may generally be relevant for his/her aligning or non-aligning actions.

Figure 56 shows that alignments are more likely to be accompanied by nods. However, it has to be stated that the frequencies of gesture types are highly dependent on the individual speaker. Regarding non-alignments, it is less clear by which specific gesture they are more likely to be accompanied.

Considering the gestural domain without the verbal and prosodic content, one could conclude that interactional participants partly rely on the other participant's gesture to determine whether or not they are allowed to continue on their current agenda. This conclusion can only be drawn if the interactional category is accepted as a fact. Of course, the verbal part of the utterance may also play an important role. A combined analysis of all modalities would be required in order to determine the relevant cues in distinguishing aligning from non-aligning actions. This is beyond the scope of the thesis.

It is interesting to note that many alignments are accompanied by gestures other than the nod, including head shakes. It is also remarkable that some non-alignments are accompanied by nods. The implication of this is that the social action is not only determined by a specific gesture, e.g. nod or head shake, but that other factors also play a role. As mentioned above, one such factor may be the verbal aspect of the utterance. Another factor may be that the social action might depend on the relationship of adjacent gestures. Similar to the approach in the acoustic analysis (Chapter 5), it can be hypothesised that alignments are characterised by gesture matching and non-alignments by gesture non-matching. The next section investigates this hypothesis.

6.1.4 Comparison of gesture adjacency

The results presented in the previous section show that the current gesture indicates the kind of action that the target IP performs. It is, however, also possible that this current gesture and its interactional function is partly determined by the prior speaker's gesture in the preceding IP. The target speaker may chose to match the prior speaker's gesture in order to perform an aligning or a non-aligning action. This section investigates whether there is a general tendency of copying the prior gesture with the current gesture. It is also investigated whether there is a difference in the strength of that tendency between the interactional categories of alignment and non-alignment. This section addresses research question RQ3b. The hypothesis is that gesture matching, i.e. the agreement between prior and current gesture, is higher for the alignment group than for the non-alignment group.

6.1.4.1 Results

Table 14 gives an overview of the distribution of prior gestures for all speakers individually and across all speakers. Figure 57 visualizes that distribution.

Table 14: Absolute and relative numbers of gestures of the prior speaker prior to the target IP.

Prior Gesture		Prior Speaker							
		A		B		C		A, B and C	
		Count	%	Count	%	Count	%	Count	%
body movement		50	24.5	9	2.9	77	26.0	136	16.7
frown		0	0.0	11	3.5	2	0.7	13	1.6
hand gesticulation		14	6.9	11	3.5	37	12.5	63	7.8
head shake		10	4.9	9	2.9	21	7.1	40	4.9
nod		58	28.4	48	15.4	92	31.1	198	24.4
none		48	23.5	199	64.0	58	19.6	305	37.6
not applicable		2	1.0	0	0.0	1	0.3	3	0.4
raised eyebrows		6	2.9	23	7.4	8	2.7	37	4.6
shrug		16	7.8	1	0.3	0	0.0	17	2.1
All	Count	204		311		296		811	
	Percentage		25.1		38.3		36.5		100.0

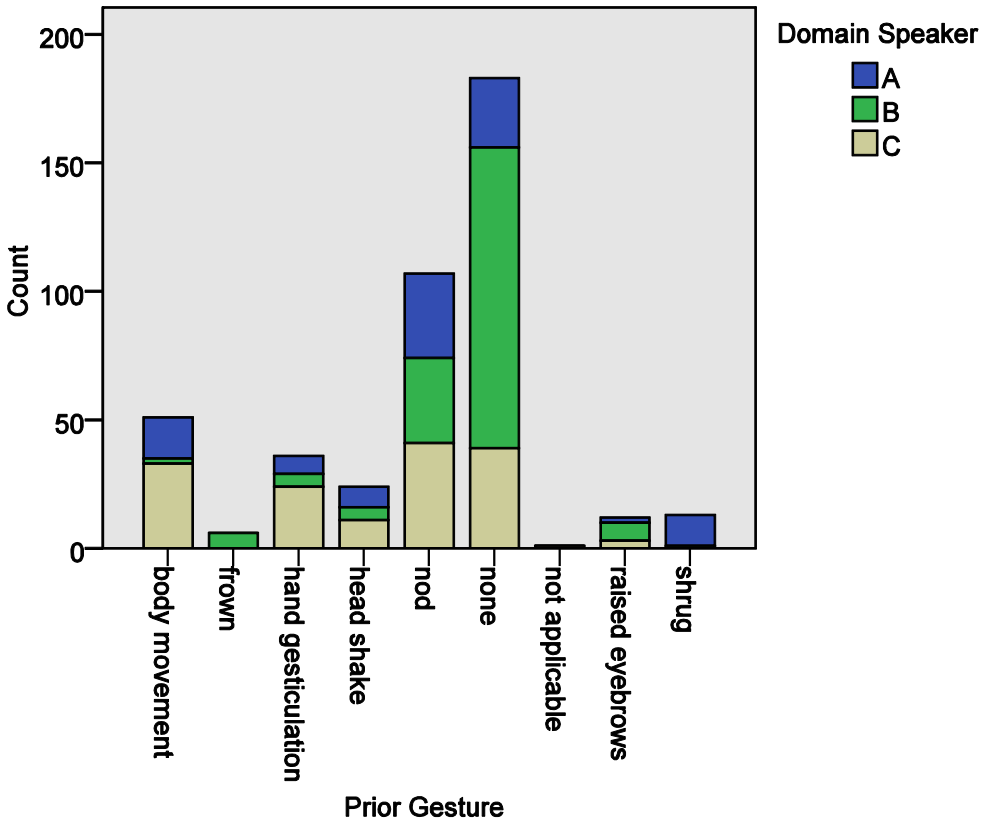


Figure 57: Gesture prior to the target IP according to the prior speaker. In most of the cases, no gesture ("none") preceded the target IP. Apart from the "none" category, nods are most frequent, followed by body movements, which are most often produced by speakers A and C. The third most frequently used gestures are hand gesticulations, followed by head shake, raised eyebrows, shrug and frown.

The speakers prior to the target IP did not use any gesture in 37.6% of the cases. There were many nods (24.4%) and body movements (16.7%) in the prior turn. The prior speaker did not frown (1.6%) or shrug very often (2.1%), but raised eyebrows (4.6%), head shakes (4.9%) and hand gesticulation (7.8%) were relatively frequent. Again, the individual gestures were not equally distributed across the three speakers. Speakers A and C perform more body movements (50 and 77) than speaker B (9), whereas speaker B seems to have a preference for frowns. Speaker C performs most hand gesticulations (37). Speaker A seems to have a preference for

shrugs (16). And speaker B produced mostly no gesture (64% of the times) compared to speaker A (23.5%) and speaker C (19.6%).

Gesture adjacency for alignments and non-alignments combined

The interrelation of succession of gestures in general (when no distinction between alignments and non-alignments is made) is summarised in Table 15.

Table 15: Prior gesture and current gesture cross-tabulation for all instances. The agreement of both gestures is 0.03 (Kappa).

Prior Gesture	Current Gesture								
	body movement	frown	hand gesticulation	head shake	nod	none	raised eyebrows	shrug	Total
body movement	8	2	0	3	55	58	10	0	136
frown	2	2	0	1	6	2	0	0	13
hand gesticulation	6	0	2	0	36	15	3	1	63
head shake	2	2	0	5	14	16	1	0	40
nod	9	14	5	3	94	65	4	4	198
none	18	15	6	32	111	106	12	3	303
raised eyebrows	2	1	1	0	16	13	3	1	37
shrug	0	0	0	1	8	6	1	1	17
Total	47	36	14	45	340	281	34	10	807

A good agreement between prior and current gestures would indicate that the target speaker generally tended to copy the prior speaker's gesture. Except for the gesture types body movement (8 out of 47), nod (94 out of 340) and none (106 out of 281), which show slight agreement, there does not seem to be a general trend to copy the prior speaker's gesture type in adjacent turn pairs. These are merely impressions based on the raw frequencies that have to be normalised, as for example nods often follow nods, which is at least partly predicted by the high occurrence of nods in our data. Here the Kappa value was chosen as described above (Section 4.3.2), which takes into account both the observed and the expected percentage of agreement. Kappa is also indicated by the variables, which are recorded on a nominal scale. The agreement between gesture types was investigated, similar to the inter-annotator agreement test (described in Section 4.3.2). The calculation of Kappa resulted in a value of 0.03. This very low value confirms that there is no general trend of gesture copying. The low agreement between gesture types indicates that performing the same gesture type is as likely as performing any one of the other gesture types in second position of an adjacent turn sequence – if the gestures are independent of the interactional categories with which they occur.

Two points can be deduced from the above. First, it is not the case that the same gesture type is selected by the target speaker more often in relation to the prior gesture than any of the other gesture types. Second, it is not the case that one of the other gesture types (not the same as the prior gesture) is selected more often than all the other gesture types (the latter would be indicated by a negative Kappa value.)

Gesture adjacency for alignments and non-alignments individually

Regarding the interactional categories of alignment (Table 16) and of non-alignment (Table 17), the agreement between prior and current gesture type is small.

The number of agreements between the type of the prior gesture and the type of the current gesture in the alignment category is generally low. The same types as in the global comparison of the gesture types (Table 15), namely "body movement" (5 out of 16), "nod" (79 out of 285) and "none" (47 out of 130) have a slightly higher agreement than the other types. A Kappa value of only 0.05 indicates low agreement, too.

Table 16: Prior gesture and current gesture cross-tabulation for alignments. The agreement of both gestures is 0.05 (Kappa).

Prior Gesture	Current Gesture								
	body movement	frown	hand gesticulation	head shake	nod	none	raised eyebrows	shrug	Total
body movement	5	0	0	2	43	27	2	0	79
frown	1	1	0	0	4	0	0	0	6
hand gesticulation	1	0	0	0	28	7	2	0	38
head shake	0	2	0	3	11	10	0	0	26
nod	1	1	1	0	79	27	0	2	111
none	8	3	2	13	99	47	3	1	176
raised eyebrows	0	0	1	0	14	8	1	1	25
shrug	0	0	0	1	7	4	0	1	13
Total	16	7	4	19	285	130	8	5	474

Table 17: Prior gesture and current gesture cross-tabulation for non-alignments. The agreement of both gestures is 0.01 (Kappa).

Prior Gesture	Current Gesture								
	body movement	frown	hand gesticulation	head shake	nod	none	raised eyebrows	shrug	Total
body movement	3	2	0	1	12	31	8	0	57
frown	1	1	0	1	2	2	0	0	7
hand gesticulation	5	0	2	0	8	8	1	1	25
head shake	2	0	0	2	3	6	1	0	14
nod	8	13	4	3	15	38	4	2	87
none	10	12	4	19	12	59	9	2	127
raised eyebrows	2	1	0	0	2	5	2	0	12
shrug	0	0	0	0	1	2	1	0	4
Total	31	29	10	26	55	151	26	5	333

The number of agreements between the prior gesture type and the current gesture type in the non-alignment category (Table 17) is even smaller (Kappa = 0.01) than it is in the case of the alignment category (Table 16).

Statistical test

This section looks at the agreement of prior and target gesture types and their relation to the interactional category. For example, if both prior and current speaker nod during their respective turns, this matching of gesture type may indicate interactional alignment. If a nod is followed by a head shake, this non-matching of gesture type may indicate interactional non-alignment.

In order to test whether the prior speaker's gesture influences the current gesture of the target speaker in performing aligning or non-aligning actions, a three way chi-square test is appropriate. But since a high proportion of the cells (36 cells; 56.3%) have an expected occurrence of gestures below 5 (see Table 15), the major requirement for conducting a valid chi-square test (i.e. a sufficient number of instances) is violated. This is also the case for Table 16 (49 cells; 76.6%) and Table 17 (46 cells; 71.9%).

The result of the chi-square test is also difficult to interpret because it depends on three parameters: the amount of independence between the variables, the sample size and the degrees of freedom. It was therefore decided to measure the "degree of association" (Pearson's phi and Cramer's V) between the two nominal variables. Because of the reasons mentioned above, the results have to be interpreted with caution.

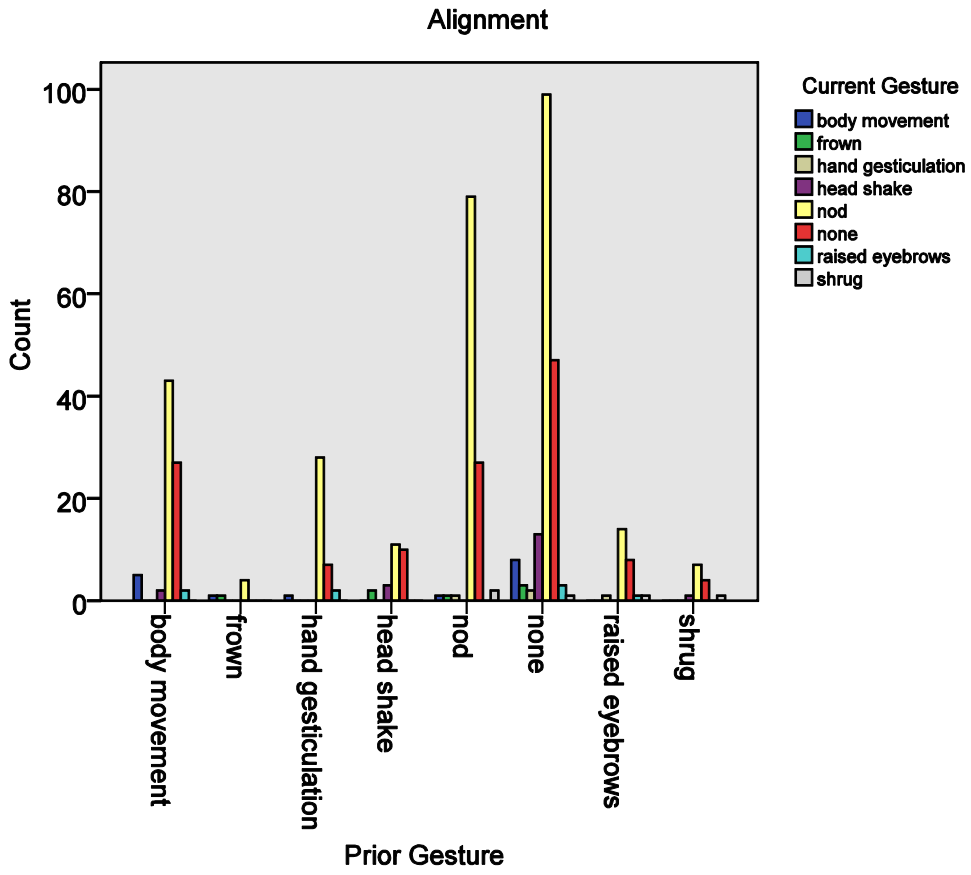


Figure 58: Frequencies of adjacent gestures (prior and current) of alignments. Most of the prior gestures seem to be predominantly followed by nods. However this is mainly due to the high occurrence of nods in general, as no association between the gestures are found to be significant.

For the current data, phi is $\phi = \sqrt{84.054/807} = 0.323$ with a significance value below 0.001, indicating that the two variables (current gesture and prior gesture) are not associated.

Cramer's V ($V = \sqrt{\chi^2/N(k-1)}$) includes one additional parameter k, corresponding to the smaller of the number of rows and columns (here both are 8). For the current data, $V = \sqrt{84.054/807 \times 7} = 0.122$ with a significance value below 0.001. This indicates that the variables are independent.

For the 474 alignments (see Figure 58), the significance level for both measures was 0.005 ($\phi = 0.407$; $V = 0.154$).

For the 333 non-alignments (see Figure 59), the significance level for both measures was 0.113 ($\phi = 0.429$; $V = 0.162$).

Again, this suggests that current gesture and prior gesture are not associated. According to the research question RQ3b, stated above, the results indicate that participants in conversation who either align or disalign with the prior speaker's agenda do not make use of gestural matching in order to distinguish the two actions.

It seems to be almost arbitrary whether the second speaker uses the same gesture type as the prior speaker or any other. However, as previously discussed, the amount of data collected is not sufficient to make strong claims. Hence, in order to rigorously test the hypothesis of gestural matching with the proposed method, it would be advisable to repeat the analysis with more data.

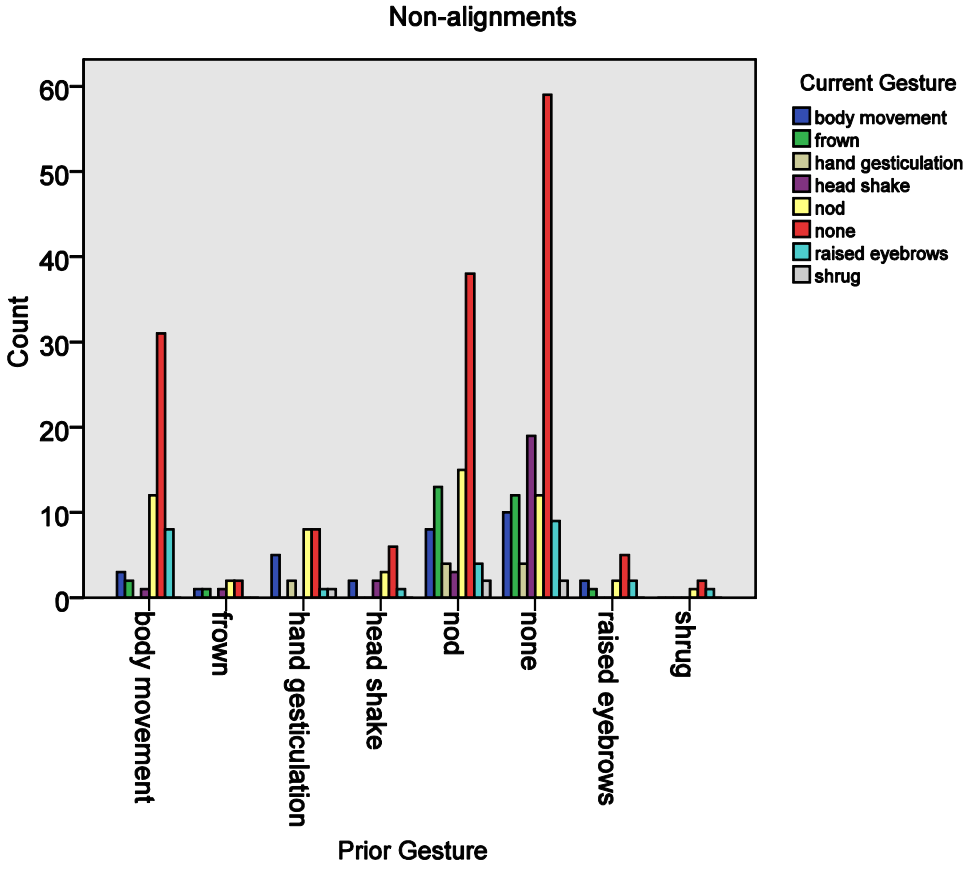


Figure 59: Frequencies of adjacent gestures (prior and current) of non-alignments. The statistics indicate no association between the two variables (prior gesture and current gesture) for the non-alignment category.

6.1.5 Discussion and summary

The influence that gestures have on communicative functions or social actions may be viewed in two ways. One view is based on the belief that the co-speech gesture itself is the relevant factor to determine the social action. The other view is based on the notion that the relative match of the current gesture to the prior gesture is the dominant factor. Both views are represented in the literature, but their findings seem to depend on the research method employed (CA vs. experimental) and, in particular, to depend on the details of the material used. For example, the interactional environments: discussion in a research meeting (current data) vs. "telling of a complaint story" (Selting, 2010) play an important role. Other confounds include the characteristics of the participants, their interpersonal relationships, the conditions of the individual meeting, and varying schemes of gesture annotation.

However, in order to come to consistent conclusions it was decided to base this study on selected research meetings taken from the well known and thoroughly documented AMI meeting corpus, in combination with an analysis of the interactional environment of short adjacent turns extracted from that corpus.

The findings reported in this chapter suggest that it is more likely that the gesture itself determines the interactional categories than the gestural match of the current gesture with the prior gesture. In other words, there are more cases where the prior speaker treats the target turn as allowing him/her to continue on the prior agenda if that turn is accompanied by a nod rather than any other gesture, for example raised eyebrows, frown or body movement.

It sometimes happened that nods that fell in the alignment category were also gestural copies of the prior speaker's nod. However, because of data sparsity, this could not be tested for statistical significance.

It should be noted that the annotation of gestures and their comparison across speakers involved a comparably rough categorisation into eight major gesture types. This trade-off in granularity was needed in order to enable inferential analyses. It might have been possible to statistically prove the gestural-matching-hypothesis if the gestures which differ in some detail were arranged into fewer (less than 8) categories. Nevertheless, according to the reasons explained in Section 3.1.3.6, it was decided to use the current set of eight gesture types.

Therefore it is not possible to compare our findings directly with the detailed qualitative gesture analyses from conversation analytic work (e.g. Selting, 2010) and purely experimental work (e.g. Holler and Wilkin, 2011a, 2011b). The current study stands in between these approaches. More data or a different organisation of gesture annotation might be necessary for an evaluation of the dependency of gestures in adjacent turns that is meaningful both in a quantitative and a qualitative sense.

The problems encountered in the current study also reveal limitations of the interdisciplinary and multimodal approach: if many disciplines and modalities are integrated in a single study, competing requirements limit the cross-section of the available data and reveal shortcomings in the different methodologies. Here, the focus was on adjacent turn pairs, a restriction imposed by the interactional analysis, which constrains the gestural analysis (and the acoustic analysis) to those turn pairs, ignoring all gestures (or acoustic phenomena) outside this narrow scope. This suggests that interdisciplinary research using multimodal approaches is difficult.

The lack of data limits the strength of the conclusions that can be drawn from the statistical tests reported here. The descriptive statistics (Kappa, Pearson's phi and Cramer's V) show that gestural matching was not used by the interactional participants in order to perform social actions such as aligning or non-aligning. The research question (RQ3b), whether gestural matching is used by participants in order to distinguish aligning from non-aligning actions has to be answered in the negative. The results are in favour of the other research question (RQ3a), whether participants use specific gestures to make a difference between aligning and non-aligning actions. It seems that alignments are accompanied by nods.

However, and finally, it may still be the case that aligning actions are performed by both, prosodic matching and by gestural matching, and that there is a bridge between the two modalities, the detail of matching being more complex (different for each modality). The difference between prosodic and gestural matching may be bound to each modality. If we put more focus on the timing relationships it may be possible to explain the differences between the two modalities.

Gestural matching can be performed simultaneously between the two speakers, while prosodic matching is usually done sequentially (except for co-productions as described by Lerner (2002)). This is possible because gestures can be performed without obscuring or masking the other speaker's gestures. If one speaker performs a gesture, another speaker's gesture can still be seen and followed visually. Spoken utterances however can mask the utterance of the interlocutor and therefore it is not possible to follow what the other speaker is saying at the same time. An exception would be if the lexical content of the other speaker's utterance is predictable enough so that it can be uttered at the same time (Lerner, 2002). This would then result in simultaneous passages. On the gestural side, such simultaneous passages are much easier to achieve and speakers do not have to wait for the other speaker to finish their utterance (gesture) in order to choose to match or non-match it.

This hypothesis could explain why we did not find evidence that gestural matches of the target speaker's turn with the prior speaker's prior turn were routinely employed to do the aligning work.

6.2 Prosodic-gestural model

Many different claims about the purpose of verbal utterances have been made. For instance, one says that there is a strong relationship between the prosodic form of an utterance and its communicative function (Levelt, 1989; Gussenhoven, 2004). Others say that there is a relationship between the prosodic form of one utterance and the prosodic form of another utterance that has just been produced by the prior speaker (Couper-Kuhlen, 1996, Szczepek Reed, 2012a). There is also evidence that two or more speakers sometimes simultaneously co-produce utterances with matching lexis and matching prosody (Lerner, 2002); this is not necessarily restricted to the verbal part of the interaction but can extend into the non-verbal, gestural modality (Selting, 2010).

A similar strong connection between prosody and gesture is suggested by Kendon (2004), whose view on the two modalities is that "Tone units are packages of speech production identified by prosodic features which correspond to units of discourse meaning. In the same way, gesture phrases are units of visible bodily action identified by kinesic features which correspond to meaningful units of action such as a pointing, a depiction, a pantomime or the enactment of a conventionalized gesture." (Kendon, 2004, chap. 7, p. 108).

The previous chapter on the acoustic analysis and the previous section have independently shown that both prosody and gesture have an influence on the social action of a second turn in an adjacent turn pair. However, the ways in which both modalities contribute are different. In the prosodic domain, the social action depends on the relative match of the target turn with the prior speaker's turn. In the gestural domain, the social action depends on the gesture of the target turn only. The question of whether the social action also depends on the gestural match could not be statistically tested, due to a lack of data.

In this section, the remaining research question addressed both prosodic similarity and gestures. It was asked how the two modalities can be combined in a sensible prosodic-gestural model. (RQ4).

From a global perspective it would make sense to combine the findings from the two modalities into a general model of interactional alignment. If the prosodic and the gestural modalities can help to predict the interactional function of a second turn, either alone or taken together, one can in principle anticipate that the combination of the two modalities would increase the quality of the prediction, i.e. the combination of both modalities would support the classification which was achieved at the interactional categorisation stage.

One may also argue that a single modality is sufficient for the other participant to decide to either continue on the prior agenda or not. Then, if the prosodic similarity is high, no nod might be necessary, or if a nod is employed, the prosody does not need to be similar.

The latter could also explain the existence of the instances which we have called false negatives and false positives. Those instances which are classified as alignments at the interactional categorisation stage, but which received a relatively low similarity score (false negatives) may be produced with a gesture supporting the initial interactional classification of alignment, i.e. a nod. Similarly, those instances which are classified as non-alignments, but which received a relatively high similarity score (false positives) may be produced with a gesture supporting the initial interactional classification as non-alignment, i.e. no nod.

6.2.1 Method

In order to evaluate the possibilities of a combined analysis, we use the results from all three analyses, the interactional analysis, the acoustic analysis and the gestural analysis.

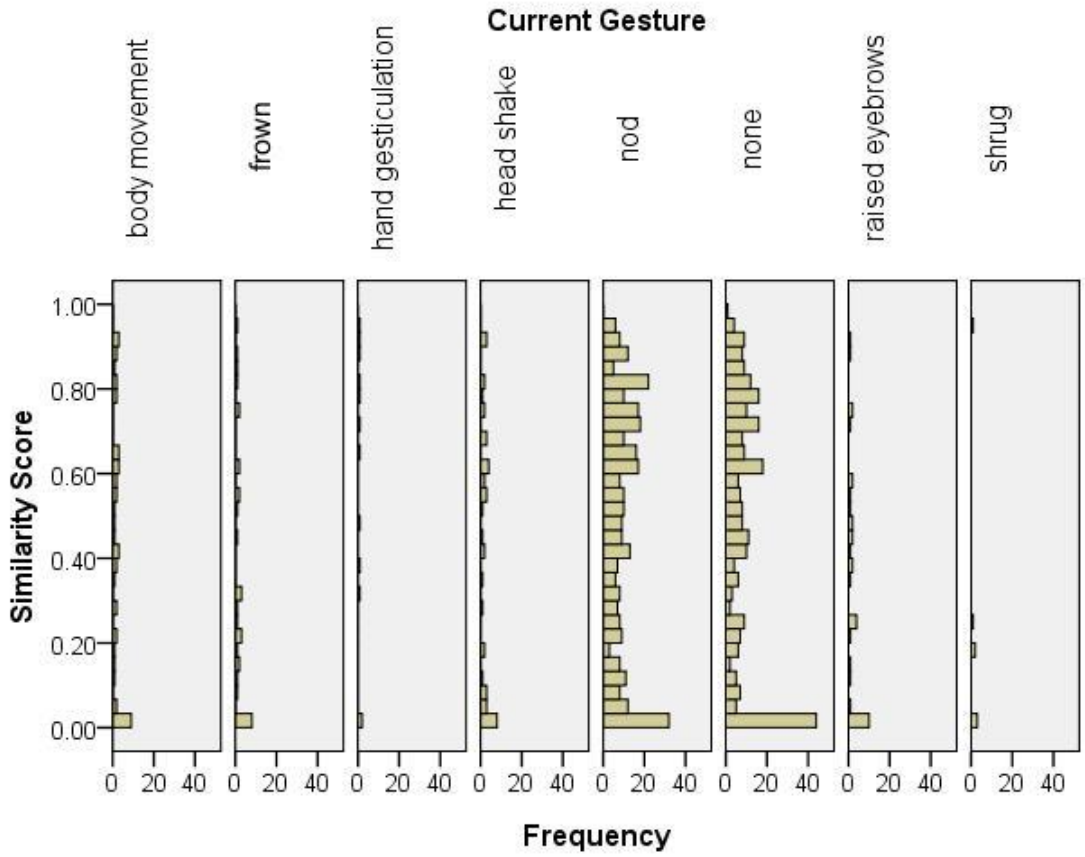


Figure 60: Distribution of similarity scores according to the target speaker's gestures (current gesture), shown in form of histograms. Some gesture types are produced by the target speaker more often (nod and none) than other gesture types. A general trend of higher or lower similarity scores for specific gesture types cannot be identified.

The interactional analysis provided us with a classification of adjacent turn pairs into two qualitative interactional categories, namely alignments and non-alignments. The acoustic analysis provided us with quantitative measures of prosodic similarity of the adjacent turn pairs. This measure indicated that there is a difference between the interactional categories according to their prosodic similarity. The gestural analysis used gesture annotations and provided us with information on the gestures with which the according actions have been performed. It is indicated that aligning actions are more likely accompanied by nods.

One possible means of analysis would be to use CA techniques to ask whether the results from the two individual analyses (acoustic and gestural) can be confirmed. This could be done by analysing the participant's orientation to adjacent turns that are prosodically similar to the prior turn, or which are performed with specific gestures. However, this is likely to be very time consuming and is not attempted here.

An alternative approach, which is pursued here, is to test with a statistical model whether the two modalities, the prosodic-acoustic modality and the gestural modality, interact in relation to the interactional categories. Such an analysis is now described.

6.2.2 Results

The distribution of the similarity score according to the target speaker's gesture is displayed as a set of histograms in Figure 60. The histograms show that the similarity score tends to spread over the whole range from 0 to 1 for all gesture types. The frequency of each gesture type varies widely, as Table 12 in Section 6.1.3 shows. It is notable that a large proportion of similarity

scores are exactly zero, which may cause difficulties in normalising the data for statistical tests (see Section 6.2.2.1).

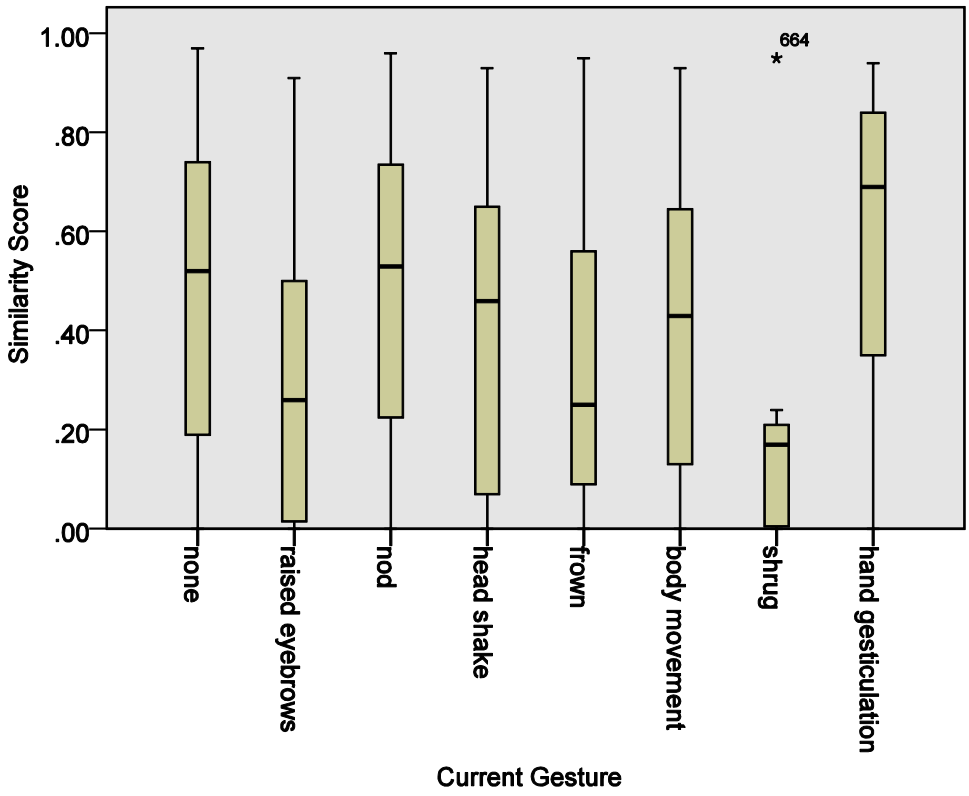


Figure 61: Distribution of similarity scores according to the target speaker's gestures (current gesture), shown in form of boxplots. The horizontal line represents the median, the length of the box corresponds to the interquartile range with the first quartile at the bottom end of the box and the upper quartile at the top end of the box. Half of the data fall within that box. The statistic analysis shows that the gesture nod is more likely to co-occur with instances of high prosodic similarity than the gesture raised eyebrows.

This data is displayed in a condensed form as boxplots in Figure 61. Raised eyebrows, frown and shrug tend to be produced more often with low similarity scores than with high similarity scores. On the contrary, hand gesticulation tends to be produced more often with high similarity scores than with low similarity scores. The other gesture types seem to be produced equally often with almost any similarity scores.

When the classification of the interactional category is included and the categories alignment and non-alignment are distinguished (see Figure 62), the overall trend seems to be that the similarity scores are split into groups of generally higher similarity scores for alignments and groups of generally lower similarity scores for non-alignments. This is expected when we take the overall differences between the two groups (see Figure 63) into account that have been found to be statistically significant, following the acoustic analysis of Chapter 5. We expect that for each gesture category, the distribution of similarity scores will tend towards high values for alignments, and low values for non-alignments. This is the case for all gesture categories, with the exception of the head shake gesture type, for which the distribution of similarity scores is very similar both for alignments and non-alignments.

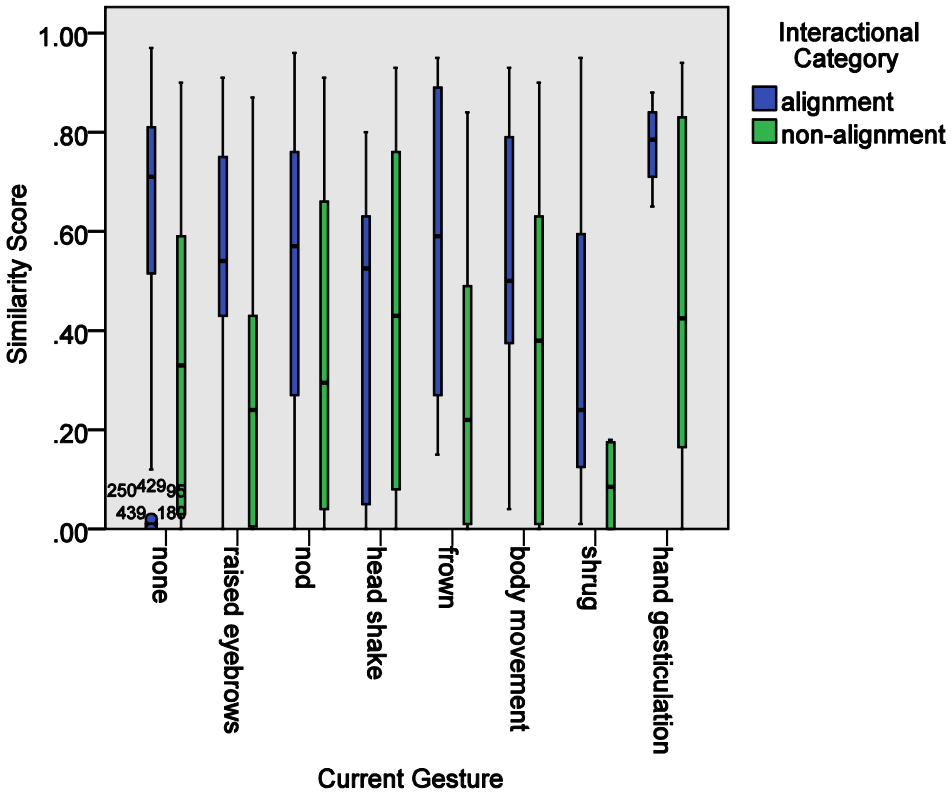


Figure 62: Distribution of similarity scores according to the target gesture types and according to the interactional categories (alignment and non-alignment), shown in form of boxplots. For each gesture type, except for the type head shake, the similarity scores tend towards high values for alignments and to low values for non-alignments.

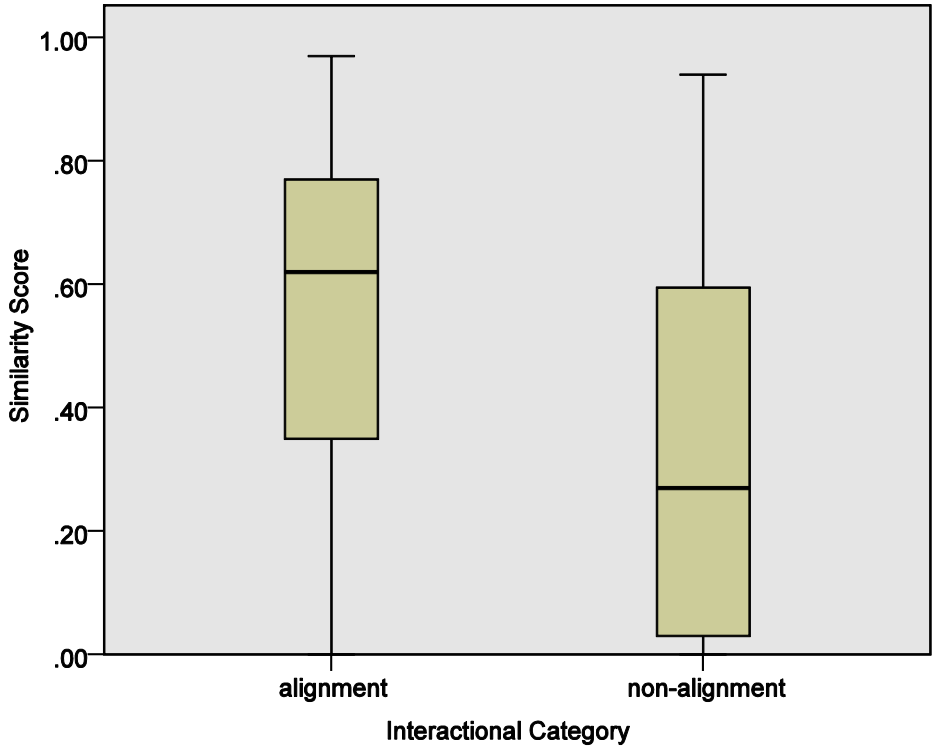


Figure 63: Distribution of similarity scores according to the interactional categories, shown in form of boxplots. Following the acoustic analysis of Chapter 5, the prosodic similarity score tends to higher values for alignments and to lower values for non-alignments.

6.2.2.1 Statistical test

A Shapiro-Wilk test of normality (Shapiro & Wilk, 1965) showed that both variables (interactional category and current gesture) are not normally distributed with the following exception: only for the gesture type hand gesticulation a normal distribution can be assumed, admitting that the overall occurrence is rather low (12 instances).

Our data therefore doesn't comply with the requirement of normality that is required for a standard analysis of variance. The acoustic analysis results are similarity values between zero and one. Data transformations were conducted in order to turn the non-normal data into normal data. Neither the square root, nor log or inverse transformation increased the distribution normality significantly. The Box-Cox transform (Osborne, 2010) was also applied in order to test if a power transform could bring the data closer to a normal distribution (all individually tested with Shapiro-Wilk tests).

The transformations of the originally observed values did not result in normal distributed data which is a requirement for parametric tests. Therefore we perform a non-parametric Kruskall-Wallis test (Kruskal & Wallis, 1952), which is based on ranks and does not require distribution normality. For each group of the independent variables (k = number of gesture types), the sum of ranks and the number of occurrences are used to calculate the mean ranks. Next, the aggregate group differences are measured:

$$group\ differences = \sum \frac{(T_g)^2}{n_g} - \frac{(T_{all})^2}{N}$$

Where T_g is the sum of ranks in the group, n_g is the number of occurrences in the group, T_{all} is the sum of ranks in all groups and N is the overall number of occurrences.

The Kruskal-Wallis procedure concludes by defining a ratio H that is defined as:

$$H = \frac{group\ differences}{N(N+1)/12}$$

As long as each group includes at least 5 samples, the sampling distribution of H is a very close approximation of the chi-square distribution for a degree of freedom $df = k - 1$. The result can then be treated as though it were a value of chi-square and referred to the sampling distribution of chi-square.

Applied to our data, the null hypothesis is that the mean ranks of the 8 groups will not substantially differ. This means that there is no difference between the distribution of similarity scores for alignments and non-alignments for gesture types. The alternative hypothesis is that there is a difference in the mean ranks of the groups.

The results of the Kruskal-Wallis test indicate that there is a statistically significant difference in the mean ranks of the similarity scores in the groups: χ^2 (7, N=766) = 22.916, p=0.002. This shows that there is a statistical interaction between the gestural category and the prosodic similarity. Because the overall test showed significant results, the pair-wise comparisons among the 8 groups were made in order to determine which of the gesture types are responsible for the interaction. The pair-wise comparisons were conducted using the Mann-Whitney *U* test, which yields results that are identical with the results from the Kruskal-Wallis test for two independent samples.

We protect for type I error, by adjusting the *a priori* alpha level divided by the number of comparisons (Bonferroni adjustment). The number of possible comparisons within 8 groups are 8*7/2=28. The adjusted alpha level is therefore 0.05/28=0.0018.

The p-values for all comparisons are shown in Table 18. The table shows that the p-value stays below this new level for only one comparison: Raised eyebrows (n=35) x nod (n=319); Z=3.243; p=0.001.

Table 18: P-values obtained from Mann-Whitney U-tests for all possible comparisons between gesture types. The only interaction with statistical significance is the one of raised eyebrows and nod. This indicates that nods are more likely than raised eyebrows to co-occur with an instance of prosodic matching. No other interaction between gesture types and prosodic similarity is statistically significant.

Gesture types / p-values	Gesture types							
	shrug	raised eyebrows	none	nod	head shake	hand gesticulation	frown	body movement
shrug		.335	.050	.033	.184	.100	.309	.124
raised eyebrows			.004	.001*	.140	.025	.796	.096
none				.680	.191	.251	.024	.309
nod					.108	.261	.011	.211
head shake						.123	.416	.830
hand gesticulation							.043	.165
frown								.208
body movement								

The result indicates that the gesture type ‘raised eyebrows’ tends to co-occur with lower similarity scores than the gesture type ‘nod’. This means that raised eyebrows are more likely to co-occur with instances of prosodic non-matching, while nods are more likely to co-occur with instances of prosodic matching.

According to the choice of the target speaker of applying specific gestures and his/her choice of prosodically matching the prior speaker’s turn, there is only a slight interaction which can be supported statistically, i.e. when the speaker raises their eyebrows, he/she has a weaker tendency to match the prior speaker’s prosody, compared to their tendency to match prosodically when nodding. For all other gesture combinations the similarity scores are not significantly different.

6.2.3 Discussion

This chapter tried to establish a link between prosodic similarity and gestural behaviour. It is proposed to take the results from the two previous chapters in order to address the question that asks how the two modalities can be combined in a sensible prosodic-gestural model (RQ4).

Therefore, we look back to these chapters. Results from chapter 5 have shown that an aligning action can be performed by prosodically matching the prior speaker’s turn. Results from chapter 6 have shown that gestural matching or non-matching did not correlate with the alignment category. Results rather suggest that specific gestures, such as nods, correlate with the alignment category.

One could conclude that the aligning action is either performed solely by the prosodic match or by specific gestures. Hence, one could assume that there probably does not exist a bridge between the two modalities.

But, there could also be an interaction between the prosodic similarity and the gesture type. The combined analysis of specific gestures and the prosodic similarity measure was tested with the Kruskal-Wallis test and the results of this comparison show that this is the case for only a very limited set of gestures. Raised eyebrows were more often used in combination with low similarity scores, while nods were more often used in combination with high similarity scores.

Addressing the above mentioned research questions, it can be stated that there might be a link between the prosodic similarity and the gestural behaviour, but in relation to the current data, this link does not seem to be very strong. It is difficult to combine the two modalities (prosody and gesture) in a sensible model, as the data is relatively sparse.

However, our combination of prosodic and gesture analysis has shown how complex multimodal interactions of conversational participants can be. Even an increase of this complexity with the sequential (consecutive) analysis of each modality can lead to interesting new findings. This was successfully demonstrated for the prosodic modality. In order to do this for the gestural modality as well, some refinements to the methodology would need to be applied, such as a larger data set or the application of motion tracking devices.

7 Discussion and conclusions

In this thesis, several research questions were asked regarding the use of gesture and prosody in talk-in-interaction. The general aim was to combine sequential interactional analysis with phonetic/prosodic and gestural analysis. One strand of research on prosody investigates prosodic features such as pitch, loudness and tempo of *individual* speech units. The emphasis is here on "individual". Another strand of research suggests that the prosodic pattern a speaker produces an utterance may also depend on the prosodic context. This is especially notable in conversation, where the speaking turns change frequently between the participants.

The analysis of conversation has led to a large knowledge base on the verbal resources which are employed by conversational participants. Some researchers found that specific prosodic characteristics are used as cues for specific social actions (Couper-Kuhlen, 2001; Gardner, 2001; Kelly and Local, 1989; Ogden, 2010). Some researchers suggested that more than a specific contour or even contrasting contours can be used to perform the same specific social actions (Walker, 2004; Kaimaki, 2010, 2011). The dependence of phonetic or prosodic characteristics across turns has been demonstrated (Couper-Kuhlen, 1996; Ogden, 2006; Lerner, 2002) and that prosodic matching is a resource to which conversational participants do indeed orient to as a relevant resource has been shown (Selting, 2010; Szczepek Reed, 2006, 2010, 2012). It suggests that more work needs to be done in order to support or reject the prosodic matching hypothesis with further instrumental investigations.

In the same way, a large knowledge base on the non-verbal resources of conversation is established. One strand of research suggests that some specific gestures are employed in order to perform specific social actions (Whitehead, 2011; Heath, 2012; McClave, 2000; Schegloff, 1987). The work gestures do, however has been shown to depend on the context in which they are used (Stivers, 2008). However only one study in CA was found that suggests that gesture matching is performed for specific conversational purposes (Selting, 2010). Work on a field which is possibly related to gesture matching is gesture mimicry (McClave, 2000; Holler & Wilkin, 2011b; McNeill, 2008).

It seems as if these hypotheses of individual prosodic or gestural patterns vs. prosodic or gestural matching stand in direct competition with each other. However, it may be the case that some social actions are performed one way and some social actions another way. The hypotheses are therefore not mutually exclusive. Nevertheless, it was found to be necessary to investigate the prosodic matching hypothesis further, as no study could be found that tested it instrumentally. Neither is there a study which contrasted specific gestures with gestural matches related to social actions in a systematic way. In this respect, the work presented in this thesis represents a novel contribution to the related fields. For interactional phonetics, it is a demonstration of how a qualitative analysis of social actions can be supported by quantitative means. For the speech science community it is a demonstration that the analysis of features can be enhanced by adding a further sequential dimension.

7.1 Answers to research questions

The first research question was related to the interactional organisation of adjacent turn pairs and the social action they perform. It was:

RQ1: What are the sequential correlates of the social actions of alignments and non-alignments?

The interactional analysis revealed that alignments were generally treated by the prior speaker as a permission to continue on his or her prior agenda. The agenda could be continued either verbally or gesturally. Non-alignments were generally treated by the prior speaker by orienting to it in a specific way. The prior agenda was not continued. A catalogue of criteria was

established using CA collecting regularities in the detail of the sequential organisation of talk that distinguished continuation of the prior speaker's agenda vs. non-continuation. Research suggesting basic interactional actions such as affiliation (Stivers, 2008; Barth-Weingarten, 2010) and alignment (Szczepek Reed, 2012a) are supported by these findings. Example transcripts were used to instruct annotators to distinguish the two social actions (alignment vs. non-alignment) on these grounds. Substantial agreement indicates that the annotators were able to recognise these details and use them in order to make a decision on the interactional category. During the annotation task the annotators merely had orthographic transcripts available.

The second set of research questions was related to the prosodic organisation of the adjacent turns and to the interactional categories alignment and non-alignment. The focus was on the dependency of prosodic patterns on the immediate prosodic context. The first out of this set was:

RQ2a: Are alignments performed with prosodic matches and are non-alignments performed with prosodic non-matches?

In the current study, the underlying interactional categories were alignments and non-alignments, following work on affiliation and disaffiliation (Stivers, 2008; Barth-Weingarten, 2010) and action continuation and non-continuation (Szczepek Reed, 2012a). For these categories it can be stated that the choice of prosodic contour depends on the prosodic context. If the second speaker aligns with the first speaker, the prosodic contour is more likely to be chosen to match the one of the previous speaker than if the second speaker does not align with the prior speaker. This supports interactional phonetics research on prosodic matching and non-matching (Couper-Kuhlen, 1996; Müller, 1996; Wells, 2010; Szczepek Reed, 2012a). A further question addressed the technical aspect of measuring prosodic similarity:

RQ2b: How can prosodic similarity be measured objectively?

The metrics that measured prosodic similarity are automatic processes analysing the acoustic signals. It is a technique which is suggested to be objective in this way. The only part of the analysis which can be influenced by subjective decisions is the collection of adjacent turns for the interactional study, however, this was done without the intention to influence the acoustic study. The interactional analysis was even a further process in between. The findings show that previous attempts to measure prosodic similarity with clean speech (Hermes, 1998b), (Rilliard, Allauzen, & de Mareüil, 2011) can be adapted to work on data from real conversations. Regarding the choice of prosodic parameters it was asked:

RQ2c: What are the prosodic parameters that are responsible for the identification of prosodic matches and non-matches?

For this study, F0 and intensity were chosen. From the positive correlation found in the comparison of the interactional categories and the prosodic similarity measures, it can be deduced that the distinction worked for the constellation of these prosodic parameters. These results and the results from an evaluation of the similarity measurement algorithms with artificial contours indicate that the choice of the prosodic parameters (F0 and intensity) was sufficient to identify prosodic matches and non-matches. Here, F0 alone was also sufficient or even superior to both parameters. It also indicates that it is possible to measure objectively the prosodic similarity of adjacent turns from two different speakers. However, several other parameters, including voice quality, duration and speech rate are candidates for inclusion in such a study (Szczepek Reed, 2006).

A third set of research questions was related to the gestural organisation of the adjacent turns and to the interactional categories alignment and non-alignment. One focused on the individual gestures used in relation to the interactional categories:

RQ3a: Are alignments and non-alignments performed with specific gestures?

From the results, it seems to be the case that the gesture in the second turn *itself* can help determine the interactional category. The other research question focused on the dependency of a gesture on the immediate gestural context:

RQ3b: Are alignments performed with gestural matches and non-alignments with gestural non-matches?

Due to lack of data it could not be tested if the interactional category was related to gestural matching or non-matching of the immediately preceding gesture of the prior speaker. Either the database needs to be increased or the number of gesture types needs to be decreased in order to be able to infer any conclusions.

Our findings support the first view as they suggest that it is more likely that the gesture itself determines the interactional category than the gestural match of the current gesture with the prior gesture. This means that the matching paradigm does not seem to be transferable from the acoustic to the gestural modality. It seems that aligning actions are not done by gestural matching, but only by prosodic matching.

The results are in favour of research that suggests that specific gestures predominantly perform the social actions of alignment (Stivers, 2008; Schegloff, 1987; Whitehead, 2011) rather than a gestural match (Selting, 2010; Lerner, 2002; McClave, 2000).

The final research question addressed both prosody and gesture. It was asked:

RQ4: How can the two modalities be combined in a sensible prosodic-gestural model?

Several possibilities could be hypothesised for achieving interactional alignment or non-alignment: First, the prosodic match could be the only responsible mechanism, irrespective of the gestural domain. Second, both domains the prosodic and the gestural domain may interact. Third, the gestural domain may be the driving factor and the prosodic domain is merely a side effect. As the acoustic analysis showed that prosodic matching helps to determine the interactional category and the gestural analysis suggests that the gesture itself, rather than a gestural match helps to determine the interactional category, a proposed model combined the individual gestures with prosodic similarity. The results reveal a statistical interaction between the prosodic similarity and the gesture used. Raised eyebrows were predominantly produced with low prosodic similarity and nods with high prosodic similarity. It has to be admitted that the available data is not sufficient enough to draw any strong conclusions on such a model. This is due mainly to the increased complexity of the interdisciplinary and multimodal approach. It would be tempting to argue that prosody and gesture have different origins, but the corpus size would need to be increased in order to evaluate this.

The main question that should be answered was how the prosodic and the gestural modalities are employed by participants of natural conversations in order to perform specific social actions. One central idea which developed from the findings is related to timing: The findings of the current study lead us to believe that different timings are an important part of the matching process. While gestures can theoretically be matched or contrasted simultaneously without obscuring each other, the acoustics cannot be produced or reproduced simultaneously (matching or non-matching) without partly masking each other. An exception is the production of choral talk in highly predictable environments (Lerner, 2002). One way of preventing the masking would be to produce the acoustics successively. Therefore, we decided to limit the scope of the current study to adjacent turn pairs. We therefore only considered the acoustic characteristics of adjacent turns, and not the characteristics of simultaneous (overlapping) turns. Similarly, the analysis of gestural properties was restricted to the adjacency of gestures – or to the gestures in second position only. This may obscure the possibility of simultaneous gestural matching, which may take place instead of consecutive gestural matching.

It should be noted that we have not dealt with specific prosodic contours cueing specific social actions. An extensive investigation of the specific prosodic characteristics would be necessary in order to give the prosodic matching a contrasting study. However, listening to the target turns reveals that all: rising and falling, high and low, loud and quiet prosodic patterns are employed for aligning purposes, as well as for non-aligning purposes. A preliminary study, which tried – but failed – to find regularities in the prosodic features of response tokens working as continuers vs. response tokens working as sequence initiations, has been presented (Gorisch, 2010), suggesting that continuers and sequence initiations do not depend on specific prosodic contours. Moreover, the positive results of the test of prosodic matching (in this thesis) show that

distinguishing alignments from non-alignments according to prosodic similarity worked, supporting the prosodic matching hypothesis (Szczepek Reed, 2012a).

7.2 Novel contributions

7.2.1 Interdisciplinary approach

At the outset of this work it was suggested that some of the most difficult questions related to human communication are more likely to be answerable with an interdisciplinary approach. One such question would be how humans organise their social interaction in order to achieve sequences of exchange that work. Another question would be how humans employ different resources (e.g. prosody or gesture) in order to achieve this. From the speech technology point of view the aim should be to apply the same processes which can be observed in human-human communication in order to make human-machine interaction as naturalistic as possible or even to understand the human speech production better.

Speech technology systems rely on data streams which are automatically accessed and processed. The data streams comprise different channels such as the audio, video and motion sensors. The process algorithms need rules in order to be able to reduce the vast amount of data into relevant analytical results for the systems' comprehension of the situation. This would finally make the systems able to interact in a naturalistic way with human users.

Although this target seems utopian, the ultimate aim of the various disciplines (phonetics, interactional phonetics, semantics, CA, etc.) is yet to find such rules. Each discipline is successful in its limited field, but is insufficient to supply the fitting algorithms. For example, phonetics investigates human speech production and perception with experimental data that includes mainly laboratory speech, but the algorithms fail once applied to real conversations. CA and interactional phonetics, which are based on real conversations, are limited by the low amount of instances that are used to explain the regularities of specific social actions in conversation, on which the rules are to be based. Therefore an interdisciplinary method was felt to be necessary.

Additionally, the new method should be standardised in order to achieve large enough collections of data so that the regularities are statistically significant. Otherwise they should not be used as rules in an algorithm. A method should also be complete, i.e. it should take into account all data which are deemed necessary to determine the regularities.

It is certainly beyond the scope of this thesis to propose a complete and mature method, but the following recommendations are proposed:

- Interactional categories should be used that are based on interactants' orientations to social actions. In this work, the categories of alignment and non-alignment were used. It was also evaluated if the sequential detail that was found to be the resource for interactional participants could also be recognised by a second analyst and used to make decisions on the interactional category using inter-rater reliability.
- In addition to prosodic analysis on the basis of individual turns, one can argue for the use of sequential prosodic-acoustic analysis. That is a prosodic analysis that takes the sequential organisation into account. It does not only search for individual prosodic properties of individual tokens, but also analyses prosodic features across speakers and across turns.
- The analysis of face-to-face meetings requires, in addition to verbal/prosodic analysis, an analysis of the non-verbal modality. Here, eight primary non-verbal gesture types were introduced and analysed according to their relation to the interactional categories. It was also analysed whether the gestural and verbal modalities interacted in relation to alignment and non-alignment.

This should not imply that these are all novel aspects of the thesis work. Introducing interactional categories has been done many times before. It is known that inter-rater reliability tests should be consistently used with subjective categorisation tasks. Gestures should be analysed if face-to-face conversations are the underlying material. The novel aspect is that all the features mentioned above are combined with respect to the interdisciplinary and multimodal approach taken in this piece of work.

7.2.2 Corpus

One premise of CA is to use recordings of naturalistic conversations. We found the research meetings of the AMI meeting corpus (http://corpus.amiproject.org/) to be a suitable database. So far, this study seems to be the first that used this corpus for CA. It contains spontaneous research meetings. Additionally, it includes individual separated audio and video channels which make a multi-modal analysis possible. The selected meetings have been pre-transcribed orthographically. Gestures were annotated by the author and can be made available to others.

7.2.3 Alignment and non-alignment

Alignments and non-alignments seem to be used in conversation systematically. In the interactional study it was attempted to map this systematic pattern onto a catalogue of criteria. If the description of such social actions should be of use to other research fields or designers of dialogue systems, it would be advisable to collect the relevant criteria which make these actions distinguishable.

From such a catalogue it might be possible to derive the underlying actions. It might be revealed that more than the two actions can be distinguished. Another possibility for an acceptable rating range than the either-or would be to quantify alignment gradually on a numerical scale rather than being a binary variable, as done here.

It was focused here on only two action types. However, more actions may be performed by the interacting participants, such as acknowledgment, agreement, disagreement, repair initiation, etc. It may well be that some of these action types are sub-types of alignments and non-alignments.

7.2.4 Measuring inter-rater reliability in CA

Researchers in CA always emphasise that the criteria for social actions always lie directly in the interaction itself (i.e. in the transcript) and can be observed and recorded by proper analysis of the participants' orientation towards each other. However, in principle, the rating of a single annotator of interactional categories is not objectively valid unless it is verified by another annotator. One analyst may recognise detail that another analyst doesn't. Subjectivity can be averaged out by taking results from more than one rater. For this reason we added a second annotator in the interactional analysis (see also Kurtić, 2012), both annotators acting as control instances with respect to the other. The validity of the two annotations was established and thereby it was shown that the criteria are indeed recognised and used. This approach has seldom been used in previous CA studies. For inter-rater reliability see Section 4.3.

7.2.5 Sequential prosodic-acoustic analysis

The prosodic-acoustic analysis is especially interesting in its ability to be readily accessible to automatic processing. This is the reason why prosodic parameters were derived and related with communicative functions. In this thesis, the prosodic resources were related with the interactional categories *alignment* and *non-alignment*. It has been shown that both correlate. F0 and intensity contours of the first and second speaker in an adjacent turn pair were used and their similarity measured. High similarity was indicative for alignments and low similarity was indicative for non-alignments. Note, that the annotators of the categories did not have access to

prosodic information during the annotation task. The description of the prosodic similarity analysis and the results can be found in Chapter 5. Mann-Whitney U-tests confirmed that alignments were more often correlated with high similarity scores and non-alignments with z-scores above 8 and p-values below 0.001 (see Section 5.4.2). As the interactional categories and the acoustic measures have a good correlation, the classification into alignments and non-alignments by the annotators can in retrospect be interpreted as a confirmation of the prosodic approach (despite a considerable number of false positives and false negatives). Therefore, the prosodic analysis has a potential to substitute the human annotator, thereby allowing automatic processing in applications such as human-machine dialogue systems.

This should not imply that acoustics are the only relevant resources for speech. Speech consists of articulatory gestures (Browman & Goldstein, 1992) that are employed, also in order to produce an acoustic signal. Sometimes, when they are visible, they can even have the potential to confuse the human speech recognition system(McGurk & MacDonald, 1976).

The acoustic analysis has still room for improvement. In this thesis, we have only analysed the fundamental frequency and intensity as prosodic parameters. Consideration may be given to extend the acoustic analysis to other parameters such as speech rate, voice quality, etc., with respect to the interactional category alignment/non-alignment.

7.2.6 Gesture

In the search for completeness of the data streams we dealt with the question whether gestures should be part of it (see Chapter 6). Are gestures used by conversational participants to express alignment and non-alignment? If yes, gestures should be included as relevant part of the data streams. In order to make the approach workable, we decided on an annotation of 9 gesture types, including body movements such as movements of the head, hand, shoulder and trunk, and facial displays such as blink, frown and raised eyebrows (see Section 3.1.3).

Two approaches towards a gesture analysis were tested. In one of them we looked at the individual gesture of the second turn in an adjacent turn sequence in relation to the interactional categories alignment and non-alignment. In the other, we compared two consecutive gestures, the gesture of the second speaker with the gesture of the first speaker, whether they reveal a gestural match. It was tested whether occurrence/non-occurrence of a gestural match correlates with alignment/non-alignment.

The single gesture analysis revealed a statistically significant correlation of specific gesture types with the interactional annotations of alignment/non-alignment (see Section 6.1.3). The analysis of gestural matches revealed no reliable results due to the non-homogeneous distribution of the gesture types (see Section 6.1.4).

As the first approach appeared to be promising, we recommend including gestures as an additional data source. There is a general trend in the literature to aim for such inclusion of gestures in dialog systems. For example head gestures are shown to be used to achieve "grounded discourse" (Morency & Darrell, 2004). However, we have to be cautious from an automation perspective: we have correlated the interactional category with single gestures, but we have not shown that a single gesture allows deducing the property alignment or non-alignment. We cannot exclude that gestures exist which are not related to alignment/non-alignment, although alignment/non-alignment are related to these gestures. In principle, the validity of a correlation does not include the validity of the reverse correlation, as becomes also apparent in a study by Stivers (2008) in which all instances of a single gesture type (nod) were annotated and analysed for their conversational action, while other sequences of that conversational action may also occur with other gestures than a nod. Therefore, the evaluation of single gestures is to be regarded as additional indication rather than a criterion.

Testing the gestural matching hypothesis, the variance of the data increases with the number of gesture types squared, when two (adjacent) gestures are taken into account. This makes it difficult to achieve a significant number of instances, unless the size of the corpus is considerably increased.

This discrepancy between the many possibilities of analysing the acoustics (recording and feature extraction) and the few possibilities of analysing the visual aspects (hand annotation of gestures) of talk shows that gesture research has some catching up to do in terms of access to objective measures of gestures (or their movement). It is necessary to find out how gestures can be integrated into the data streams for further processing. Video recordings and motion tracking with sensors are two ways which can help to extract gestures.

The following section is a short excursion to present an idea of tracking movements of participants which are engaged in conversation. The idea has been realised in a short project with another student (Shaabi Mohammed) and colleagues (Emina Kurtić, Guy Brown, Bill Wells) at the University of Sheffield (2010).

7.3 Motion tracking

Orthographic transcripts are crucial instruments in the objective analysis of social interactions. If acoustic features such as prosodic parameters need references, they can be measured according to the acoustic signals and can be used objectively to support a researcher's claims. Video recordings need to be transcribed in order to be of use in qualitative analyses of face-to-face interactions. Compared with the transcription of audio data, this can be rather difficult: When is a head movement a nod when a head shake or when a shrug? How salient is the motion? In which direction does it start or end? How often does the head move upwards and downwards? An attempt for annotating such detail can be found in Whitehead (2011) Synchronisation of multiple channels is equally crucial in multimodal analyses: When does a nod have its strongest acceleration – on the first or the second syllable? When does the nod start – before or after a verbal start?

Figure 64: Miniature chips with accelerometers and gyros are attached to the holder of the headset microphone at the side of the head. The SunSPOT (in front of the participant) transmits the recorded motions (50/s) wirelessly to a laptop.

(photo by the author, reproduced with permission of the persons shown, names known to the author)

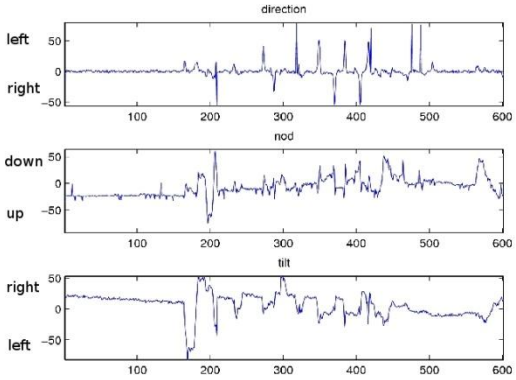


Figure 65; Sample graphs of three motion features (direction, nod and tilt) calculated from accelerometer and gyroscope measures (50/s). Positive and negative values indicate rotation for direction (left, right), tilt for nod (downward, upward) and tilt (left right).

In order to obtain data collections from recordings with a level of detail that enables us to answer the questions above, we exploited recent developments in motion tracking technology. There are commercial motion tracking facilities on the market. However they generally have shortcomings: they are bulky and sometimes they need clear visibility between small devices like reflectors or locators which are attached to parts of the body and the recording video camera. Additionally image processing software is necessary to track the trajectories of these

small devices frame by frame. These techniques are rather high priced exceeding the average budget of a university institute. Recent low-cost hardware can now achieve good results, as was demonstrated in a six week summer project at the University of Sheffield. We used head-mounted devices attached to SunSPOTs (see Figure 64) in order to record acceleration and orientation from meeting participants, which allowed the transmission of data wirelessly to a laptop computer.

We used a chip with three accelerometers and two gyroscopes in order to capture three features. The rotation around the vertical axis was used to indicate direction (left-right). The angle of tilt in the direction forward-backward (or downward-upward) was used to indicate nodding. The angle of tilt in the direction left-right was used to indicate the head tilt in either direction which could be used to indicate a shrug.

Ten minutes of natural conversations of three participants were recorded on video using this motion tracking system. Figure 65 illustrates the signals obtained from a short stretch of recordings for one of the participants.

The short duration of the project (six weeks) was not sufficient to get to a stage where longer natural conversations could be recorded. However, we have demonstrated that recording of motion signals is possible even under a limited budget. Video recordings and recordings of motion could be used as the basis for gesture recognition. In combination with the audio signals, the information obtained from these sensors could be used to finally substitute hand annotation of gestures or at least make hand annotation more meaningful. It could also provide objective measures for characterising gesture in a similar way as the features extracted from the acoustics for characterising prosody. A similar approach to gesture matching would be possible. This is interesting future work.

7.4 Suggestions for future study

7.4.1 Change in the direction of analysis

Szczepek Reed points out that "connecting back to prior talk becomes a necessity every time participants produce next turns" (2012a, p. 18), and we have achieved this with objective measures of prosodic similarity. However, our basic methodology was different to the one employed in Szczepek Reed's studies (2006, 2012a), where prosodically matching or non-matching turns were extracted from impressionistic transcripts which contained prosodic annotations, and then analysed interactionally. In order to avoid circularity, we concentrated on the interactional analysis of transition relevance places first and then fed the categorised examples into the acoustic-prosodic analysis afterwards. This step pattern has also been used by Sikveland (2012), who analysed the interaction first and analysed the phonetics and the prosody afterwards.

An alternative approach is to combine Szczepek Reed's method with our automatic prosodic analysis. Instead of manually selecting prosodic matches and non-matches, the prosodic similarity or dissimilarity could be computed across whole conversations. Afterwards, all the stretches which are identified as very similar or dissimilar could be used for interactional analysis. This is an interesting direction for future research.

7.4.2 Novel corpus

One premise of CA is that basic social actions performed by participants of a naturally occurring conversation can be identified irrespective of the setting in which the conversation takes place (at the work place, in a casual meeting, at the dinner table, in court, etc.) and irrespective of the conversational participants taking part (young, old, male, female, etc.). Whatever the participants do in the specific setting is the norm (for them) – and if it deviates

from that norm, the participants show orientation towards the trouble source indicating what was not normal.

However, it is not possible to extrapolate the findings from one set of data and generalise that they are also true for other conversational settings and for different participants. It is only through analysis of different participants and settings that individual trends can be tested to be observable in general. This is also part of the CA methodology. Therefore it would be necessary to validate the findings from the current study by applying the proposed steps of analysis to a novel corpus with different speakers and see if similar regularities can be found.

As the current study identified several weaknesses of the corpus used, there is space for improvement. For example the audio recordings include sources of noise (breathing noise) and crosstalk on individual headset microphones. With lapel microphones the influence of breathing noise can be avoided by a certain degree, but then the interferences due to crosstalk increases. Dedicated sound separation techniques need to be applied to overcome these discrepancies. On the recordings of visual information, individual cameras make it possible to annotate the movements of the participants. This is prone to error as conventionalised transcription conventions for gesture which are approvable for inter-rater reliability too, are not yet standardized. An automation of the gesture annotation process would help to make their analysis more objective and accessible.

7.5 Implications

In this study, we have presented evidence that a speaker's choice of pitch contour is locally managed. This finding has wide-ranging implications. We have already mentioned those for the modelling of prosodic features in applications such as automatic speech recognition and dialogue systems. Instead of focussing on specific prosodic contours for specific interactions with the user, it might be useful to adjust the prosodic contours to the user's prosodic contours depending on the envisaged social action. Similarly, the user's prosodic characteristics could be monitored in response to the system's prosodic output, in order to evaluate the social action the user might have envisaged.

Another implication is for the understanding of how children develop their use of prosody (Wells, 2010). Prosodic matching and non-matching might be the first resources children use in order to perform most basic social actions. If a lexicon of prosodic contours connected with specific communicative meanings develops, this might happen at a later stage – or never.

It might also have implications for the understanding of an atypical development such as the immediate and delayed echolalia found in cases of low-functioning autism (Local & Wootton, 1995) or other developmental disorders or a blind person (Fay & Coleman, 1977).

It may also have implications for how intonation is taught to second language learners (Szczepek Reed, 2012b). It seems to be necessary to take the influence of the prosodic context in conversational talk into account, especially the role of prosody for interactional alignment.

Our findings may also have implications for research in neuroscience looking at the chameleon effect (Chartrand & Bargh, 1999) or the activity of brain regions in the production and perception of speech in relation to real conversation (Scott, McGettigan, & Eisner, 2009). Prosody has also been brought in connection with research on empathy (Aziz-Zadeh, Sheng, & Gheytanchi, 2010). Aziz-Zadeh et al. indicate that areas in the brain, which are important for prosody production and prosody perception may be utilised for "aspects of social communication and social understanding, such as aspects of empathy..." (p. 1) How far prosodic matching is related to empathetic behaviour is an interesting avenue for future research.

8 References

Aimetti, G. (2011). *A Computational Model of Early Language Acquisition: a data-driven approach inspired by the empiricist view of cognitive development.* PhD thesis: University of Sheffield.

Aimetti, G., Moore, R. K., & ten Bosch, L. (2010). Discovering an optimal set of minimally contrasting acoustic speech units: a point of focus for whole-word pattern mtching. *Interspeech.* Makuhari, Japan.

Aimetti, G., Moore, R. K., ten Bosch, L., Räsänen, O., & Laine, U. (2009). Discovering keywords from cross-modal input: ecological vs. engineering methods for enhancing acoustic repetitions. *Interspeech.* Brighton.

AMI. (2008, June 17). *AMI Corpus Meeting IDs Explained.* Retrieved April 09, 2012, from AMI Corpus: http://corpus.amiproject.org/documentations/

AMI. (2010, June 15). *AMI Annotations.* Retrieved April 09, 2012, from AMI project: http://corpus.amiproject.org/documentations/annotations

AMIDA. (2010, February 04). *Final public report.* Retrieved April 09, 2012, from AMI project: http://www.amiproject.org/newsletter/issue-20

Atkinson, J. M., & Heritage, J. (1984). *Structures of Social Action: Studies in Conversation Analysis.* Cambridge: Cambridge University Press.

Aziz-Zadeh, L., Sheng, T., & Gheytanchi, A. (2010). Common premotor regions for the perception and production of prosody and correlations with empathy and prosodic ability. *PLoS ONE, 5*(1), 1-8. doi:10.1371/journal.pone.0008759

Ball, M. J., & Local, J. (1996). Current developments in transcription. In M. J. Ball, & M. Duckworth, *Advances in Clinical Phonetics* (pp. 51-89). Amsterdam/Philadelphia: John Benjamins.

Barth-Weingarten, D. (2011). Double sayings of German JA—more observations on their phonetic form and alignment function. *Research on Language and Social Interaction, 44*(2), 157-185.

Bavelas, J. B., & Gerwing, J. (2007). Conversational hand gestures and facial displays in face-to-face dialogue. In K. Fiedler, *Social communication* (pp. 283-308). New York: Psychology Press.

Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes, 15*, 469-489.

Beckman, M. E., & Hirschberg, J. (online). *The ToBI Annotation Conventions.* Retrieved April 09, 2012, from Ohio State University - ToBI: http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html

Birdwhistell, R. L. (1970). *Kinesics and Context.* Philadelphia: University of Pennsylvania Press.

Boersma, P., & Weenink, D. (2012). Praat: doing phonetics by computer. *[Computer Program]*. Amsterdam. Retrieved from http://www.fon.hum.uva.nl/praat/

Bolinger, D. (1965). Pitch accent and sentence rhythm. In I. Abe, & T. Kanekiyo, *Forms of English: Accent, Morpheme, Order* (pp. 139-180). Tokyo: Hokuou.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica, 49*(3-4), 155-180.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation Journal, 41*(2), 181-190.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology, 76*(6), 893-910.

Chovil, N. (1991). Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction, 25*, 163-194.

Chovil, N. (2005). Measuring conversational facial displays. In V. Manusov, *The sourcebook of nonverbal measures: Going beyond words* (pp. 173-188). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 163-194.

Cooke, M. P. (1993). *Modelling auditory processing and organisation.* Cambridge: Cambridge University Press.

Couper-Kuhlen, E. (1996). The prosody of repetition: On quoting and mimicry. In E. Couper-Kuhlen, & M. Selting, *Prosody in conversation: Interactional studies* (pp. 366-405). Cambridge: Cambridge University Press.

Couper-Kuhlen, E. (2001). Interactional prosody: High onsets in reasons-for-the-call turns. *Language in Society, 30*, 29-53.

Couper-Kuhlen, E. (2004). Prosody and sequence organization in English conversation: The case of new beginnings. In E. Couper-Kuhlen, & C. E. Ford, *Sound patterns in interaction* (pp. 335-376). Amsterdam: John Benjamins.

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

Doughty, M. (2001). Consideration of three types of spontaneous eyeblink activity in normal humans: during reading and video display terminal use, in primary gaze, and while in conversation. *Optometry & Vision Science, 78*(10), 712-725.

Drew, P. (1987). Po-faced receipts of teases. *Linguistics, 25*(1), 219-253.

Drew, P. (1997). 'Open' class repair initiators in response to sequential sources of trouble in conversation. *Journal of Pragmatics, 28*(1), 69-101.

Drew, P. (2004). Conversation Analysis. In K. L. Fitch, & R. E. Sanders, *Handbook of Language and Social Interaction* (pp. 71-102). Mahawa, New Jersey, and London: Lawrence Erlbaum Associates.

Drew, P., & Walker, T. (2009). Going too far: Complaining, escalating and disaffiliation. *Journal of Pragmatics, 41*, 2400-2414.

Efron, D. (1941). *Gesture and Environment.* Morningside Heights, NY: King's Crown Press.

Ekman, P., & Friesen, W. (1969). The repertoire of nonverbal behavior: Categories, origins, and coding. *Semiotica, 1*, 49-98.

Ewan, W. G. (1975). Explaining the intrinsic pitch of vowels. *Journal of the Acoustical Society of America*, 40.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874.

Fay, W. H., & Coleman, R. O. (1977). A human sound transducer/reproducer: Temporal capabilities of a profoundly echolalic child. *Brain and Language, 4*(3), 396-402.

French, P., & Local, J. (1983). Turn-competitive incomings. *Journal of Pragmatics, 7*(1), 17-38.

Fry, D. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America, 27*, 765-769.

Fry, D. (1958). Experiments in the perception of stress. *Language and Speech, 1*, 205-213.

Gardner, R. (1997). The conversation object mm: A weak and variable acknowledging token. *Research on Language and Social Interaction, 30*(2), 131-156.

Gardner, R. (2001). *When listeners talk: Response tokens and listener stance.* Amsterdam; Philadelphia: John Benjamins Publishing Company.

Geller, T. (2012). Talking to machines. *Communications of the ACM*, 14-16.

Geluykens, R. (1987). Intonation and speech act type: An experimental approach to rising intonation in queclaratives. *Journal of Pragmatics, 11*(4), 483-494.

Goodwin, C. (1980). Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning. *Social Inquiry, 50*(3-4), 272-302.

Goodwin, C., & Goodwin, M. H. (1987). Concurrent operations on talk: Notes on the interactive organization of assessments. *IPRA Papers in Pragmatics, 1*(1), 1-54.

Goodwin, M. H. (1980). Processes of mutual monitoring implicated in the production of description sequences. *Sociological Inquiry, 50*, 303-317.

Gorisch, J. (2010). Prosodic matching of response tokens in conversational speech. *Colloquium of the British Association of Academic Phoneticians (BAAP2010).* London, UK.

Gorisch, J., Wells, B., & Brown, G. J. (2012). Pitch contour matching and interactional alignment across turns: An acoustic investigation. *Language and Speech, 55*(1), 57-76.

Gravano, A. (2009). *Turn-taking and affirmative cue words in task-oriented dialogue.* PhD thesis, NY: Columbia University.

Gravano, A., & Hirschberg, J. (2009). Back-channel-inviting cues in task-oriented dialogue. *Interspeech*, (pp. 1019-1022). Brighton, UK.

Gravano, A., Benus, S., Chávez, H., Hirschberg, J., & Wilcox, L. (2007). On the role of context and prosody in the interpretation of okay. *ACL*, (pp. 800-807). Prague, Czech Republic.

Grice, M., Reyelt, M., Benzmüller, R., Mayer, J., & Batliner, A. (1996). Consistency in transcription and labeling of German intonation with GToBI. *ICSLP*, (pp. 1716-1719). Philadelphia, USA.

Gussenhoven, C. (2004). *The phonology of tone and intonation.* Cambridge: Cambridge University Press.

Gut, U., & Bayerl, P. S. (2004). Measuring the reliability of manual annotations of speech corpora. *Speech Prosody*, (pp. 565-568). Nara, Japan.

Habets, B., Kita, S., Shao, Z., Özyürek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech-gesture integration during comprehension. *Journal of Cognitive Neuroscience, 23*(8), 1845-1854.

Hamon, C., Mouline, E., & Charmentier, F. (1989). A diphone synthesis system based on time-domain prosodic modifications of speech. *ICASSP*, (pp. 238-241). Glasgow, Scotland.

Harris, J. D. (1947). The effect of sensation level upon pitch discrimination in a continuous thermal noise mask. *Journal of the Acoustical Society of America, 19*(4), 733-733.

Hawkins, S. (2011). Does phonetic detail guide situation-specific speech recognition? *International Congress of Phonetic Sciences*, (pp. 9-12). Hongkong, China.

Heath, C. (1992). Gesture's discrete tasks: Multiple relevancies in visual conduct and in the contextualization of language. In P. Auer, & A. di Luzio, *The contextualization of language* (pp. 101-127). Amsterdam: John Benjamins.

Heldner, M., Edlund, J., & Hirschberg, J. (2010). Pitch similarity in the vicinity of backchannels. *Interspeech*, (pp. 3054-3057). Makuhari, Japan.

Hermes, D. J. (1998a). Auditory and visual similarity of pitch contours. *Journal of Speech,*

Language, and Hearing Research, 41(1), 63-72.

Hermes, D. J. (1998b). Measuring the perceptual similarity of pitch contours. *Journal of Speech, Language, and Hearing Research, 41*(1), 73-82.

Holle, H., Obermeier, C., Schmidt-Kassow, M., Friederici, A. D., Ward, J., & Gunter, T. C. (2012). Gesture facilitates the syntactic analysis of speech. *Frontiers in Psychology, 3*(74), 1-12.

Holler, J., & Wilkin, K. (2011a). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior, 35*(2), 133-135.

Holler, J., & Wilkin, K. (2011b). An experimental investigation of how addressee feedback affects co-speech gestures accompanying speaker's responses. *Journal of Pragmatics, 43*, 3522-3536.

Holmes, J., & Holmes, W. (2001). *Speech Synthesis and Recognition* (2 ed.). London and New York: Taylor & Francis, Inc.

Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction, 28*, 171-183.

Jefferson, G. (1984). Notes on a systematic deployment of the acknowledgement tokens 'yeah' and 'mmhm'. *Papers in Linguistics, 17*(2), 197-216.

Joh, A., & Hosoma, H. (2010). Simultaneous gestural matching in multi-party conversation highlights the difference between participants. *International Conference on Conversation Analysis (ICCA10).* Mannheim, Germany.

Kaimaki, M. (2010). Tunes in free variation and sequentially determined pitch alignment: Evidence from interactional organisation. *Journal of Greek Linguistics, 10*, 213-250.

Kaimaki, M. (2011). Sequentially determined function of pitch contours: the case of English news receipts. *York Papers in Linguistics 2, 11*, 50-73.

Kelly, J., & Local, J. (1989). *Doing Phonology: observing, recording, interpreting.* Manchester: Manchester University Press.

Kendon, A. (1982). The study of gesture: some observations on its history. *Recherche Semiotique/Semiotic Inquiry, 2*(1), 25-62.

Kendon, A. (2004). *Gesture: Visible Action as Utterance.* Cambridge: Cambridge University Press.

Kimbara, I. (2006). On gestural mimicry. *Gesture, 6*(1), 39-61.

Kimbara, I. (2008). Gesture form convergence in joint description. *Journal of Nonverbal Behavior, 32*(2), 123-131.

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes, 24*(2), 145-167.

Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and Cognitive Processes, 22*(8), 1212-1236.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America, 118*(2), 1038-1054.

Kousidis, S., Dorran, D., McDonnell, C., & Coyle, E. (2009). Convergence in human dialogues time series analysis of acoustic feature. *SPECOM2009.* St. Petersburg, Russian Federation.

Kousidis, S., Dorran, D., Wang, Y., Vaughan, B., Cullen, C., Campbell, D., McDonnell, C., & Coyle, E. (2008). Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues. *Interspeech.* Brisbane, Australia.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411-433.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47,* 583-621.

Kurtić, E. (2012). *Overlapping talk and turn competition in multi-party conversations.* PhD thesis: University of Sheffield.

Kurtić, E., Brown, G. J., & Wells, B. (2009). Fundamental frequency height as a resource for the management of overlap in talk-in-interaction. In D. Barth-Weingarten, N. Dehé, & A. Wichmann, *Where prosody meets pragmatics (Studies in Pragmatics 8)* (pp. 183-204). Bringley: Emerald.

Kurtić, E., Brown, G. J., & Wells, B. (in press). Resources for turn competition in overlapping talk. *Speech Communication*.

Lakin, J. L., Jefferis, V. E., Cheng, C. M., & Chartrand, T. L. (2003). The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior, 27*(3), 145-162.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Lerner, G. H. (2002). Turn-sharing: the choral co-production of talk-in-interaction. In C. Ford, B. Fox, & S. Thompson, *The Language of Turn and Sequence* (pp. 225-256). Oxford: Oxford University Press.

Lerner, G. H. (2003). Selecting next speaker: The context-sensitive operation of a context-free organization. *Language in Society, 32*(2), 177-201.

Lerner, G. H. (2004). On the place of linguistic resources in the organization of talk-in-interaction: Grammar as action in prompting a speaker to elaborate. *Research on Language and Social Interaction, 37*(2), 151-184.

Levelt, W. J. (1989). *Speaking.* Cambridge, MA.: The MIT Press.

Local, J. (1992). Continuing and restarting. In P. Auer, & A. di Luzio, *The contextualization of language* (pp. 273-296). Amsterdam: John Benjamins.

Local, J. (2003). Phonetics and talk-in-interaction. *International Congress of Phonetic Sciences,* (pp. 115-118). Barcelona, Spain.

Local, J. (2007). Phonetic detail and the organisation of talk-in-interaction. *International Congress of Phonetic Sciences.* Saarbrücken, Germany.

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

Local, J., & Walker, G. (2005). Methodological imperatives for investigating the phonetic organization and phonological structures of spontaneous speech. *Phonetica, 62,* 120-130.

Local, J., & Wootton, T. (1995). Interactional and phonetic aspects of immediate echolalia in autism: a case study. *Clinical Linguistics & Phonetics, 9*(2), 155-194.

Local, J., Kelly, J., & Wells, B. (1986). Towards a phonology of conversation: turn-taking in urban Tyneside speech. *Journal of Linguistics, 22,* 411-437.

Loehr, D. P. (2004). *Gesture and Intonation.* Washington, DC: Georgetown University.

Mandelbaum, J. (1990). Beyond mundane reason: Conversation analysis and context. *Research on Language and Social Interaction, 24,* 333-350.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*(1), 50-60.

McClave, E. (2000). Linguistic functions of head movements in the context of speech. *Journal*

of Pragmatics, 32(7), 855-878.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* Chicago: University of Chicago Press.

McNeill, D. (2008). Gesture of power and power of gesture. *Berlin Ritual-Conference.* Berlin: Dec. 5. 2008, E. Fischer-Lichte and C. Wulf (editors).

Meyer, C. (2010). Gestenforschung als Praxeologie: Zu Jürgen Streecks mikroethnolographischer Theorie der Gestik. *Geschprächsforschung, 11*, 208-230.

Mol, H., & Uhlenbeck, E. (1956). The linguistic relevance of intensity in stress. *Lingua, 5*, 205-213.

Moore, J., Kronenthal, M., & Ashby, S. (2005). *Guidelines for AMI speech transcriptions.* IDIAP, Switzerland; University of Edinburgh. Edinburgh: AMI project. Retrieved from http://www.amiproject.org/

Moore, R. K. (2007). PRESENCE: A human-inspired architecture for speech-based human-machine interaction. *IEEE Transactions on Computers, 56*(9), 1176-1188.

Morency, L.-P., & Darrell, T. (2004). From conversational tooltips to grounded discourse: head pose tracking in interactive dialog systems. *Proceedings of the 6th international conference on multimodal interfaces (ICMI '04)*, (pp. 32-37). State College, PA, USA.

Müller, F. E. (1996). Affiliating and disaffiliating with continuers: Prosodic aspects of recipiency. In E. Couper-Kuhlen, & M. Selting, *Prosody in conversation: Interactional studies* (pp. 131-176). Cambridge: Cambridge University Press.

Nachtigall, C., & Wirtz, M. (2004). *Wahrscheinlichkeitsrechnung und Inferenzstatistik, Statisitische Methoden für Psychologen, Teil 2* (3 ed.). Weinheim and Munich: Juventa.

Neiberg, D., Salvi, G., & Gustavson, J. (2013). Semi-supervised methods for exploring the acoustics of simple productive feedback. *Speech Communication, 55*(3), 451-469.

Neuhoff, J. G., & McBeath, M. K. (1996). The doppler illusion: The influence of dynamic intensity change on perceived pitch. *Journal of Experimental Psychology: Human Perception and Performance, 22*(4), 970-985.

Nilsenová, M., Swerts, M., Houteoen, V., & Dittrich, H. (2009). Pitch adaption in different age groups: boundary tones versus global pitch. *Interspeech*, (pp. 1015-1018). Brighton, UK.

Ogden, R. (2006). Phonetics and social action in agreements and disagreements. *Journal of Pragmatics, 38*, 1752-1775.

Ogden, R. (2010). Prosodic constructions in making complaints. In D. Barth-Weingarten, E. Reber, & M. Selting, *Prosody in Interaction* (pp. 81-103). Amsterdam: John Benjamins.

Ogden, R. (2012). The phonetics of talk in interaction – introduction to the special issue. *Language and Speech, 55*(1), 3-11.

Orton, J. (2007). Gesture in modern language teaching and learning. *Babel, 42*(2), 12-18.

Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation (electronic journal), 15*(12), 1-9. Retrieved from http://pareonline.net/pdf/v15n12.pdf

Park, S. H., Goo, J. M., & Jo, C.-H. (2004). Receiver Operating Characteristic (ROC) curve: Practical review for radiologists. *Korean Journal of Radiology, 5*(1), 11-18.

Parril, F., & Kimbara, I. (2006). Seeing and hearing double: The influence of mimicry in speech and gesture on observers. *Journal of Nonverbal Behavior, 30*(4), 157-166.

Paukner, A., Suomi, S. J., Visalberghi, E., & Ferrari, P. F. (2009). Capuchin Monkeys display

affiliation towards humans who imitate them. *Science, 325*, 880-883.

Pitrelli, J., Beckman, M., & Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. *ICSLP*, (pp. 123-126). Yokohama, Japan.

Plack, C. J. (2005). *The Sense of Hearing.* Mahwah, NJ: Lawrence Erlbaum Associates.

Pomerantz, A. M. (1978). Compliment response: Notes on the cooperation of multiple constraints. In J. N. Schenkein, *Studies in the Organisation of Conversational Interaction* (pp. 79-112). New York: Academic Press.

Pöppel, E. (2009). Pre-semantically defined temporal windows for cognitive processing. *Philosophical Transactions of the Royal Society B, 364*, 1887-1896.

Rabiner, L. R., & Juang, B. (1993). *Fundamentals of speech recognition.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Riek, L., Paul, P. D., & Robinson, P. (2010). When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. *Journal of Multimodal User Interfaces, 3*, 99-108.

Rilliard, A., Allauzen, A., & de Mareüil, P. B. (2011). Using dynamic time warping to compute prosodic similarity measures. *Interspeech.* Florence, Italy.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696-735.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing, 26*(1), 43-49.

Scheflen, A. E. (1964). The significance of posture in communication systems. *Psychiatry, 27*, 316-331.

Scheflen, A. E. (1968). Human communication: Behavioral programs and their integration in interaction. *Behavioral Sciences, 13*, 44-55.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In D. Tannen, *Georgetown University Roundtable on Languages and Linguistics (1981) Analyzing Discourse: Text and Talk* (pp. 71-93). Georgetown: Georgetown University Press.

Schegloff, E. A. (1984). On some gestures' relation to talk. In J. M. Atkinson, & J. Heritage, *Structures of Social Action: Studies in Conversation Analysis* (pp. 266-296). Cambridge: Cambridge University Press.

Schegloff, E. A. (1987). Analyzing single episodes of interaction: An exercise in conversation analysis. *Social Psychology Quarterly, 50*(2), 101-114.

Schegloff, E. A. (1996). Turn organization: One intersection of grammar and interaction. In E. Ochs, E. A. Schegloff, & S. Thompson, *Interaction and Grammar* (pp. 52-133). Cambridge: Cambridge University Press.

Schegloff, E. A. (2007). *Sequence organization in interaction. A primer in Conversation Analysis* (Vol. 1). Cambridge: Cambridge University Press.

Schmidt, T. (2003). Visualising linguistic annotation as interlinear text. *Arbeiten zur Mehrsprachigkeit, Folge B*, 1 ff. Retrieved from http://www.exmaralda.org/files/Visualising-final.pdf

Schmidt, T. (2010). Another extension of the stylesheet metaphor -- Visualising multi-layer annotations as musical scores. In A. Witt, & D. Metzing, *Linguistic modelling of information and Markup Languages* (pp. 23-44). Dordrecht: Springer.

Scott, S. K., McGettigan, C., & Eisner, F. (2009). A little more conversation, a little less action -- candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience, 10*, 295-302.

Selting, M. (2010). Affectivity in conversational storytelling: An analysis of display of anger or indignation in complaint stories. *Pragmatics, 20*(2), 229-277.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrica, 52*(3-4), 591-611.

Sikveland, R. O. (2012). Negotiating towards a next turn: phonetic resources for 'doing the same'. *Journal of Language and Speech, 55*(1), 77-98.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). TOBI: A standard for labelling English prosody. *ICSLP*, (pp. 867-870). Banff, Alberta, Canada.

Sluijter, A. (1995). *Phonetic Correlates of Stress and Accent.* The Hague: Holland Academic Graphics.

So, W. C., Kita, S., & Goldin-Meadow, S. (2009). Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Science, 33*, 115-125.

Steensig, J., & Drew, P. (2008). Introduction: Questioning and affiliation/ disaffiliation in interaction. *Discourse Studies, 10*(5), 5-15.

Steensig, J., & Larsen, T. (2008). Affiliative and disaffiliative uses of you say x questions. *Discourse Studies, 10*(5), 113-133.

Stivers, T. (2008). Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Research on Language and Social Interaction, 41*(1), 31-57.

Streeck, J. (2009). *Gesturecraft. The Manu-facture of Meaning.* Amsterdam: John Benjamins.

Szczepek Reed, B. (2006). *Prosodic orientation in English conversation.* Houndmills, Hampshire, UK: Palgrave MacMillan.

Szczepek Reed, B. (2009). Prosodic orientation: A practice for sequence organization in broadcast telephone openings. *Journal of Pragmatics, 41*(6), 1223 – 1247.

Szczepek Reed, B. (2010a). Prosody and alignment: A sequential perspective. *Cultural Studies of Science Education*, 859-867.

Szczepek Reed, B. (2010b). Intonation phrases in natural conversation: A participants' category? In D. Barth-Weingarten, E. Reber, & M. Selting, *Prosody in Interaction* (pp. 191-212). Amsterdam: John Benjamins.

Szczepek Reed, B. (2012a). Beyond the Particular: Prosody and the Coordination of Action. *Journal of Language and Speech, 55*(1), 13-34.

Szczepek Reed, B. (2012b). Prosody in conversation: Implications for teaching English pronunciation. In J. Romero-Trillo, *Pragmatics, prosody and English language teaching* (pp. 147-168). London: Springer.

Szczepek Reed, B., & Raymond, G. (forthcoming). *Units of Talk – Units of Action.* Amsterdam: John Benjamins.

Tannen, D. (1989). *Talking voices: repetition, dialogue, and imagery in conversational discourse.* Cambridge: Cambridge University Press.

Tape, T. G. (2003, December 02). *The area under an ROC curve.* Retrieved from Interpreting diagnostic tests: http://gim.unmc.edu/dxtests/roc3.htm

Ten Have, P. (1990). Methodological Issues in Conversation Analysis. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 27*, 23-51.

Walker, G. (2004). *The phonetic design of turn endings, beginnings, and continuations in conversation (PhD thesis).* York: University of York.

Walker, G. (2012). Coordination and interpretation of vocal and visible resources: 'trail-off' conjunctions. *Journal of Language and Speech, 55*(1), 141-163.

Wang, D., & Brown, G. J. (2006). *Computational auditory scene analysis: Principles, Algorithms, and Applications.* Piscataway, NJ: IEEE Press/Wiley-Interscience.

Ward, N. (2000). Issues in the transcription of English conversational grunts. *1st SIGdial Workshop at ACL.* HongKong, China.

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

Wells, B. (2010). Tonal repetition and tonal contrast in English carer-child interaction. In D. Barth-Weingarten, E. Reber, & M. Selting, *Prosody in Interaction* (pp. 243-262). Amsterdam: John Benjamins.

Wells, J. C. (2006). *English Intonation: an introduction.* Cambridge: Cambridge University Press.

Whitehead, K. (2011). Some uses of head nods in 'third position' in talk-in-interaction. *Gesture, 11*(2), 103-122.

Wichmann, A. (2011, unpublished). Prosody in context — the effect of sequential relationships between speaker turns.

Wightman, C. W. (2002). ToBI or Not ToBI? *Speech Prosody*, (pp. 25-29). Aix-en-Provence.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. *LREC.* Genoa, Italy.

Wood, J. M. (2007, October 03). Understanding and computing Cohen's Kappa: A tutorial. *WebPsychEmpiricist*. Retrieved from http://works.bepress.com/james_wood/22/

Wundt, W. (1973). *The Language of Gestures* (4 ed., Vol. 1). (J. S. Thayer, C. M. Greenleaf, & M. D. Silberman, Trans.) The Hague: Mouton.

Yngve, V. (1970). On getting a word in edgewise. *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pp. 567-577.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103*, 347-387.

Appendix A. Transcription conventions

Two types of transcripts are used in this thesis. First, in Extract 1 and Extract 2, the traditional style is adapted with the following symbols:

hh Audible outbreath; number of characters indicating the duration in 0.1 s steps.

.hh Audible inbreath.

: Lengthening of preceding speech, for example a long hesitation "u::m".

(0.3) Pause in seconds.

(.) Micro pause (less than 0.1 s).

[Beginning of overlapping speech.

* Truncated speech at starts or ends of words, for example "ha*" or "*tion".

Second, in the transcripts in Chapter 4 in the style of musical scores used the following symbols:

hh° Audible outbreath

°hh Audible inbreath

: Lengthening

(0.3) Pause in seconds

(.) Micro Pause

(-) Short pause

* Truncated speech

_ Spelling; for example "G_D_F_" pronounced as "gee dee ef".

((sound: ...)) Sounds such as laughter, clicks, etc. without standard orthographic symbols.

Appendix B. Training for second annotator

Annotation Training for Interactional Categories

Table of Contents

1 Two basic categories: agenda alignment vs. non-alignment

2 Transcripts

The underlying material are orthographic transcripts which include the verbal content of the speaker's turns as well as the speakers' gestures.

The outline is a bit different to the traditional CA style. It is more like a musical score with the speakers and their gestures standing for the instruments. Starting turns do not automatically start a new line. They are interwoven with all other speaker's turns. All speakers' turns continue this way until the end of the line (bunch of lines) and are wrapped into the next line (bunch of lines). See examples below.

3 Interactional environment

We are interested in the oral turns of one speaker which follows an oral turn of another speaker. We call these turns 'first turn' and 'second turn'. For the 'second turn' we search its interactional category. In order to find what action it performs in the ongoing interaction, we have to look at its environment. This environment is both oral and gestural. And both give us hints on the action. So we annotate actions.

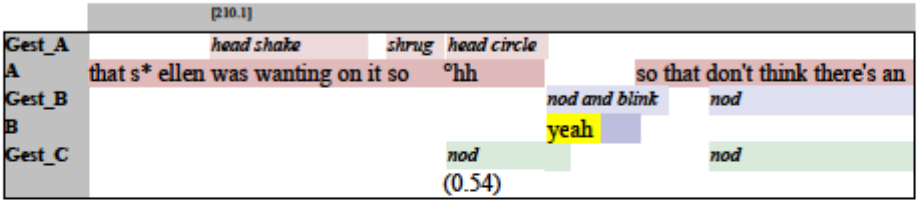
4 Sequences

4.1 Extract “calibration”

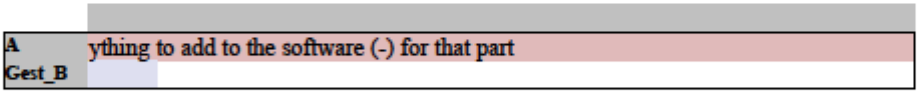
[1]

Gest_A	*body movement*	*body movement*
A	it loo* it looks like it was doing all the calibration	and drift correction things

[2]



[3]

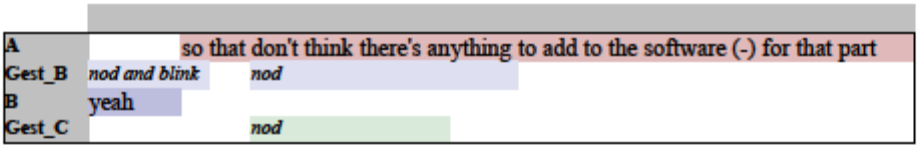


– First speaker (A) continues after the target turn "yeah" of second speaker (B) on the first speaker's prior agenda.

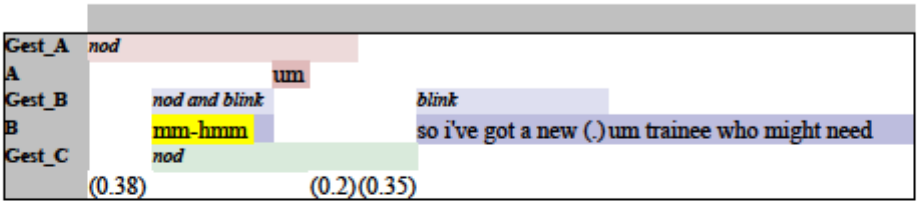
=> Alignment

4.2 Extract "software"

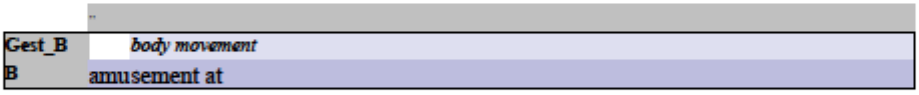
[1]



[2]



[3]



– First speaker (A) continues with a hesitation
– First speaker continues same gesture as before and during the target turn.
– Target turn "mm-hmm" is followed by a pause
– After the pause the second speaker continues

=> Alignment

4.3 Extract "eye link"

[1]

Gest_B						*raised*
B	((sound: laugh))		okay	yeah	i mean	
Gest_C		*hand gesticulation*	*body movement*			
C	uh yeah so we've moved over and (.)		on to the other eye link so uh			
		(0.07)			(0.54)	

[2]

Gest_B	*eyebrows*		*raised eyebrows*		
B			m*	my view on this	was (0.14) that um: the
Gest_C		*raised eyebrows*			
C		'cause she's	running at the moment		

[3]

B	need for people to back up their

Instance "yeah":

- Similar to Extract "software":
  - Target turn ("yeah") followed by short pause
  - After pause: second speaker continues ("I mean")
  - however: no gesture from first speaker (C)

=> Alignment

Instance "I mean":

- First speaker does not continue on prior agenda

=> Non-alignment

Instance "my view on this":

- overlaps first speaker's turn
- Second speaker (B) continues on different agenda than first speaker's agenda

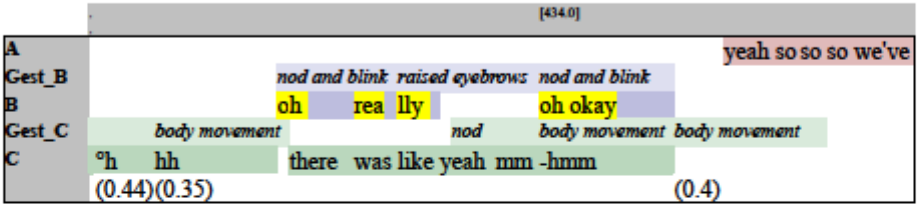
=> Non-alignment

4.4 Extract "more like an eye"

[1]

Gest_C	*body movement*	*hand gesticulation*
C	ellen wanted something to look more like an eye (.) so that you knew that was the eye	

[2]



[3]

Instance "oh really":

- Target turn overlaps with continuation of first speaker
- Third turn ("yeah") of the first speaker is not on the first speaker's agenda ("there was like")
- Gesture of the first speaker (C) in third turn (after target) is different to the gesture before the target

=> Non-alignment

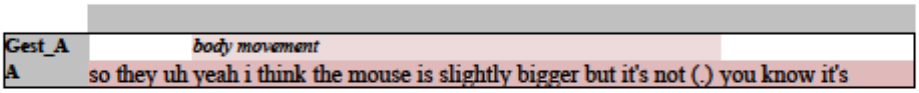
Instance "oh okay":

- Target turn overlaps with continuation of first speaker
- Pause after target turn
- Gesture of the first speaker (C) after the target is the same as during the target
- Third speaker (A) takes the floor after the pause

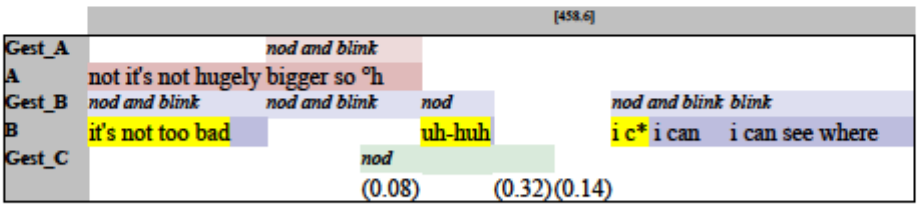
=> Alignment

4.5 Extract "slightly bigger"

[1]



[2]



[3]

Gest_B	
B	that's useful

Instance "it's not too bad":

- Overlaps with prior speaker
- Prior speaker continues on prior agenda

=> Alignment

Instance "uh-huh":

- Pause after the target
- Target speaker continues after the pause on his/her own agenda

=> Alignment

Instance "I c*":

- New agenda starts (different to first speaker's agenda)

=> Non-alignment

4.6 Extract "free package"

[1]

Gest_A	*nod*		*raised eyebrows*
A	i'd b* i'd be very surprised if there wasn't a free package for doing it (.)		u:m (0.31) (0.55)

[2]

Gest_A		*nod*	*frown*	
A		i ha* i have a feeling that the: um amaya the ((w_)) three ((c_))		
Gest_B	*nod and blink*			
B	uh-huh			yep

[3]

Gest_A	*nod*	*nod*
A	browser	does it h° have a feeling that that can actually edit
Gest_B	*nod and blink*	*nod and blink*
B	mm-hmm	

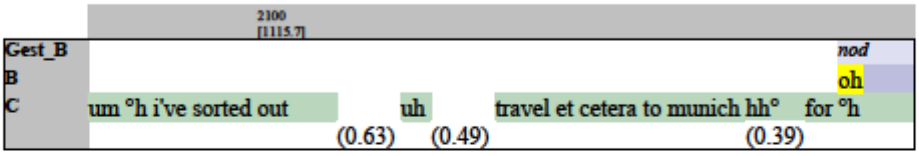
Both instances "uh-huh" and "yep mm-hmm":

- first speaker (A) continues on the prior agenda

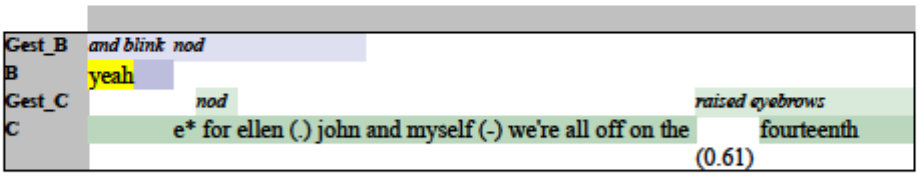
=> Alignment

4.7 Extract "travel et cetera"

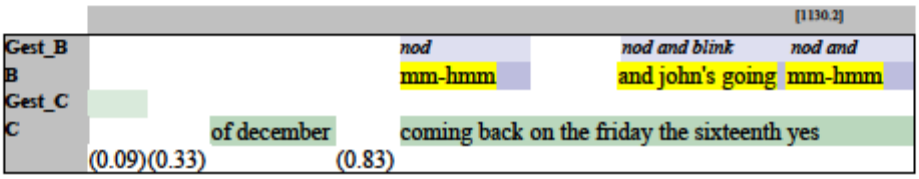
[1]



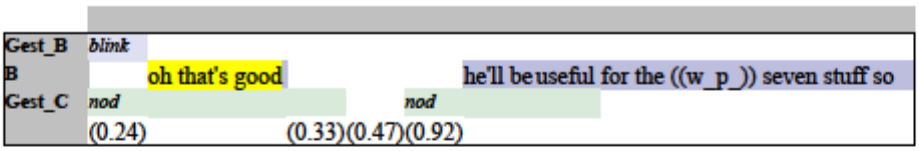
[2]



[3]



[4]



Instances "oh yeah" and "mm-hmm":

- First speaker (C) continues on his/her agenda

=> Alignment

Instance "and john's going":

- Third turn ("yes" after target) does not continue the prior agenda of the same speaker

=> Non-alignment

Second instance "mm-hmm":

- Short pause after the target
- Same speaker (target speaker B) continues on different agenda than first speaker's agenda
- First speaker's gesture after the target turn continues the gesture prior and during the target

=> Alignment or Non-alignment?

4.8 Summary

There is a nice way to test whether some turn is a continuation on the prior agenda or not. If first and the third turn (of the same speaker) are glued together and still make sense, the agenda is continued. If it becomes obvious that something is missing between the two parts which have been glued together, then the agenda is not continued in the third turn.

Examples:

"... ellen was wanting on it so that don't think there's an ..."

"... to add to the software to that part um ..."

"... on to the other eye link so uh 'cause she's ..." *

"... there was like yeah mm-hmm ..." *

On the gestural side we can also test the "agenda characteristics". If the first speaker continues after the target turn with the gesture he/she was using before or during the target turn, the gestural agenda is continued. If the gesture after the target turn is not continued, the gestural agenda is not continued.

A problem arises when the oral and gestural agendas contradict each other. Then it comes to a conflict and other details might give a clue for the classification as in Extract "polygons".

4.9 Extract "polygons"

[1]

Gest_A	*nod*					
Gest_B		*nod and blink*				
B		oh that's good				
Gest_C	*body movement*			*nod*		
C	the polygons now don't	overlap		°hh	um °h marloes discovered something	
				(0.62)		

[2]

C	strange about the orientation of parallelograms from when

- Pause after the target
- It us unclear whether the third turn (after the target) is continuing the agenda of the first turn or if something new is initiated.
- The gesture of the first speaker after the target (nod) is different to the gesture before the (body movement) target

=> Alignment or non-alignment?

5 Task

In the next days I will prepare the collection of transcripts. Then I send you that material and instructions.

5.1 Classification

For each instant, analyse whether the target turn is an 'alignment' or a 'non-alignment'.

The main criterion is whether the prior speaker continues after the target on his/her agenda.

The following principles might help for most of the cases.

If the agenda is continued, the target was an alignment.

If the first speaker does not continue on his/her agenda, but orients to the target in a specific way, the target is a non-alignment.

If the first speaker does not orally, but gesturally continue, the target was an alignment

If the first speaker does not take the turn at all after the target, it depends on the target speaker's continuation:

If there is a short break after the target and before the target speaker continues on a new agenda, the target was an alignment.

If there is no short break between the target and the same speaker's continuation, the target was a non-alignment.

5.2 Saving (in preparation)

Save the category 'alignment' or 'non-alignment' in one table column, and the confidence level for the decision on a three point scale from 0 to 2.

"confident" = 2

"fairly sure" = 1

"just guessing" = 0

Appendix C. Complete list of instances

Table 19: List of all instances of adjacent turn pairs (target turn and prior turn) collected from the AMI meetings EN2009b, EN2009c and EN2009d (meeting B, C or D). The interactional category after conversation analysis (CA) is indicated with A for alignment and N-A for non-alignment. The prosodic similarity (Accumulative quality score) between the two turns using the dynamic time warping technique (DTW) is indicated in column DTW_{Sim}.

Meeting/ Start	Target Turn	Prior Turn	CA	DTW_{Sim}
B1	okay	we've moved over and	A	
B2	eye link	okay	A	
B3	yeah	eye link	A	0.95
B4	I mean	so uh	N-A	0.14
B5	my view on this	cause she's running at the moment	N-A	0.44
B13	mm-hmm	to lab users	A	
B14	she did admit it	I mean I guess	N-A	0.72
B30	hmm	more	A	0.65
B30	no	we can do really	A	0.8
B42	yeah there was a mailing list	should know about it now	N-A	0.52
B43	mm-hmm	before	A	0.42
B45	Frank	right	N-A	0.03
B52	right okay	so the	A	0.79
B54	yeah	updated	A	0.22
B55	uh-huh	updated	A	0.62
B55	so everyone on that list	uh-huh	N-A	0.66
B57	mm-hmm	yeah	A	0.74
B63	right	so	A	0.8
B63	yes	so	A	0.72
B64	and we don't know	yes	N-A	0.34
B69	yeah I mean	right	N-A	0.09
B71	yeah I mean	apart from	N-A	0.18
B74	right	hard drive check	A	0.29
B78	yeah	but something	A	0.43
B79	uh-huh	in the first place	A	0.36
B80	yeah yeah	so	A	0.53
B84	uh-huh	and hope	A	0.23
B88	yeah	that'll be it	A	0.61
B89	so how long does this	that'll be it	N-A	0
B96	day 'n a bit	two days	A	0.94
B97	day and a bit	day and a bit	A	0.88
B97	yeah	last week	A	0.56
B97	by the time everything	day and a bit	A	0.87
B99	yeah	by the time everything	N-A	0.8
B99	uh-huh	by the time everything	A	0.73
B99	and then I'll be	yeah yeah	N-A	0.93
B104	yeah I mean	today	N-A	0.22
B114	yep	isn't t	A	
B115	mm-hmm	yep	A	
B115	and I saw a b*	yep	N-A	

B121	um the eye-link	you mean from	A	0.95
B122	I	was that Joe or	N-A	0.19
B131	so anyway	um	N-A	0.79
B137	oh no I mean	model builder	N-A	0
B139	uh well that wa*	the eye-link software	N-A	0.65
B142	so it's unrelated	but yeah	N-A	0.81
B144	no	that	A	
B145	right	no	A	
B145	okay good	it's unrelated	A	0.11
B154	yeah	experiment	A	0.85
B160	uh he's	put his schedule back	N-A	0.77
B162	he's done it	done his	N-A	0.71
B163	okay	he's done it	A	0.71
B163	yes	got his	A	0.69
B164	and he backs up	okay	N-A	0.85
B193	yeah	collaborating between the two machines	A	0.69
B202	oh so you need to borrow	do a proper	N-A	0.01
B205	yeah	and stuff to	A	0
B206	yeah	right	A	0.66
B207	but i* it looks	and	N-A	0
B212	yeah	so	A	0.72
B215	mm-hmm	for that part	A	0.05
B216	so I've	um	N-A	0
B231	spare head	be here	N-A	0.06
B237	yeah	around	A	0.8
B250	don't think so	did you guys meet her	A	0.15
B263	well I did	be willing to	N-A	0.01
B269	oh who is that	evening it was	N-A	0.72
B270	Heidi	oh who is that	A	0.8
B271	oh yeah	Heidi	N-A	0
B271	Heidi	Heidi	A	0.73
B273	oh yeah she's good	and	N-A	0.26
B274	yeah mm-hmm	oh yeah she's good	A	0.2
B274	you guys	yeah mm-hmm	N-A	0
B279	yeah	and gesture	A	0.27
B283	and autism	and autism	A	0.24
B292	she's been volunteered	Nynke to volunteer	N-A	0.77
B305	yeah	people use	A	0.43
B306	yeah so try an*	so	N-A	0.25
B325	mm-hmm	audio signal as well	A	
B334	mm-hmm	slightly easier	A	0.66
B334	uh yeah	slightly easier	N-A	0.52
B341	oops	the soundproof box	N-A	0
B360	right	so	A	0.35
B364	okay	control panel	A	0.12
B365	so we want one	control panel	N-A	0
B368	yeah	Matlab	A	0.05
B382	okay	really clear	A	0.6

B383	perfect	exactly where the beep is	A	0.74
B399	yeah yeah	bank the script to some place	A	0.53
B400	yeah	tell us how to run it	A	
B400	uh-huh	yeah	A	
B401	cause we won't know	no problem	N-A	0.55
B408	what else	sounds good	N-A	0.9
B416	yeah	and the mouse	A	0.01
B433	oh really	that was the eye	N-A	0.47
B434	yeah	oh really	A	0.27
B434	oh okay	yeah	N-A	0.6
B444	uh-huh	a mouse	N-A	0
B444	okay	a mouse	N-A	0
B445	very nice	so	N-A	0.66
B450	uh no the	what's underneath	N-A	0.54
B452	*riginal one mm-hmm	as the original blobs	A	0.87
B457	it's not too bad	you know	A	0.79
B458	uh-huh	bigger so is	A	0.47
B459	I c*	bigger so is	N-A	0
B461	yeah	useful	A	0.57
B461	and and if	useful	N-A	0
B464	just change it	reduce the size	A	0.76
B465	mm-hmm	of whatever	A	0
B466	as that's not too hard	mm-hmm	A	0.01
B476	yeah	purposes	A	0.37
B476	yeah	purposes	A	0.51
B487	mm-hmm	jiggle around right	A	0.15
B488	but large	but	N-A	0.77
B512	yes he's	the model builder	N-A	0.34
B524	oh that's good	polygons now don't	A	0.62
B548	okay	problem or not	A	0.34
B556	really	model builder	N-A	0.45
B557	yeah	really	N-A	0.21
B562	just the parallelogram	with rotation	N-A	0.45
B564	yeah just the parallelogram	just the parallelogram	A	0.66
B576	right	something else	A	0.57
B576	okay so	something else	N-A	0.62
B584	Joe	pass it to	A	0.71
B584	yep	Joe	A	
B585	okay	yep	A	
B585	that makes sense	yep	A	
B595	mm-hmm	you know	N-A	0.32
B618	uh-huh yeah yeah	the vertex points	A	
B626	yeah	an existing set of	A	0.74
B626	so you need to	an existing set of	N-A	0.66
B627	so you need to draw them in the first place	so you need to draw them in something	A	0.77
B628	yeah I suggested	so you need to draw them in the first place	N-A	0.78
B631	yeah yeah	thing to do or	N-A	0.43

B632	yes	no	A	0.76
B636	yep	a comra	A	0.3
B642	yeah	part	A	0.29
B642	so this	part	N-A	0.01
B651	okay	put SVG	A	0.7
B652	but they may need help	put SVG	N-A	0.78
B655	yeah yeah	I suppose	A	
B656	so I'll I'll	I suppose	N-A	
B663	Microsoft though	I men I kno* I kno* I know	N-A	0.74
B663	Microsoft though	pink package	N-A	0.22
B666	(yeah) but I know	well	N-A	0.16
B669	they are the paid	but obviously they're like	A	0.64
B675	uh-huh	for doing it	A	0.37
B680	yep mm-hmm	the W three C	A	0.83
B684	okay	SVG	A	0.83
B684	so I wouldn't	files	N-A	0
B699	yeah yeah	for something	A	0.11
B700	you know	yeah yeah	A	0.82
B700	let them	yeah yeah	N-A	0.88
B706	yeah	what'll do it	A	0.04
B706	yeah fair enough	what'll do it	A	0.08
B706	it's not	what'll do it	N-A	0.06
B715	uh-huh yeah yeah	or something	A	0.63
B718	oh so	it's a	N-A	0.25
B733	no	right	A	0
B737	um right	it should be all right	A	0.78
B737	it doesn't matter	it should be all right	N-A	0.44
B742	no it sh*	concave in	A	0.25
B743	well	funny ways	A	0.49
B743	uh hopefully	funny ways	N-A	0.24
B748	yeah	it does have sort of	A	0.45
B749	yeah	and things	A	0.78
B755	and you tried it	just yeah	N-A	0.61
B756	it just has to be	but it just	A	0.52
B756	yeah yeah yeah	and you tried it	A	0.82
B756	yeah yeah yeah	but it just	A	
B756	it just has to be	and you tried it	A	
B766	would we want	I have to check that but	N-A	0.84
B770	a part that crosses over itself	by self-intersection you mean	A	
B773	you mean like	over itself	N-A	0.64
B779	yeah yeah yeah	as parts	A	0.43
B785	by self-intersection	for rotation	N-A	0.27
B790	yeah	something like that	A	0.79
B791	yeah	yeah	A	0.88
B798	yeah	yeah	A	0
B812	no	now	A	0.63
B814	yeah	important enough	A	0.92
B815	yeah	yeah	A	0.88

B815	so for now	yeah	N-A	0.17
B833	yep mm-hmm	to the start point	A	0.3
B835	mm-hmm	so	A	0
B842	uh-huh	should be all right	A	0
B850	yeah	an SVG	A	0.46
B853	but it they have	any problems	N-A	0
B858	yeah yeah	worth while	A	0.49
B865	yeah	so	A	0.79
B873	well	from ex-fig	N-A	0.25
B887	that I wasn't gonna say	or something yes	N-A	0.52
B901	yeah	so	A	0.89
B903	but in my	yeah	N-A	0.1
B907	that should be fine	that should be	A	0.68
B915	nn-hnn	can't find anything	A	0.15
B927	nope	or	A	0.59
B928	uh I think	happy at the moment	N-A	0.93
B942	oh they've done the pilots	or English speaking	N-A	0.26
B943	yeah	the pilots	N-A	0.26
B943	oh good	yeah	A	0.1
B947	dumped the data	other than	A	0.75
B947	dumped the data	dumped the data	A	0.94
B948	where is it dumped	dumped the data	N-A	0.34
B958	well that's	onto the	N-A	0
B963	yeah	some way	A	0.55
B963	um	yeah	N-A	0.18
B976	you know	further is	N-A	0.09
B994	the CVS	the CVS	A	0.8
B1001	oh so you're	easy and	N-A	0.01
B1004	nope	CVS before	A	0.5
B1005	okay	nope	A	0.8
B1009	yeah yeah yeah	and no	A	0.89
B1012	oh so far	that's reassuring	N-A	0.39
B1013	yeah	yeah	A	0.75
B1014	cause he's sitting on	yeah	N-A	0.54
B1015	yes	desk	A	0.12
B1023	yeah	piece of software	A	0.72
B1035	mm-hmm	that's it	A	
B1045	yeah	but	A	
B1057	mm-hmm	it's all	A	0
B1061	that's it we'll never see him here again	once I've got the logins	A	0.94
B1068	not yet	as well	A	0.42
B1083	no	supplying	A	0.55
B1084	moment	right	A	0.7
B1102	uh yeah	or do they	N-A	0.65
B1113	uh-huh	at the moment	A	0.39
B1120	oh yeah	to Munich	A	0.49
B1128	mm-hmm	of December	A	0.53

B1129	and John's going	on the Friday	N-A	0.42
B1130	yes	and John's going	A	0.19
B1131	ah that's good	yes	N-A	0.61
B1147	both right	both	A	0.87
B1161	uh o* one	again	N-A	0.01
B1170	yeah	stuff	A	0.6
B1179	right	on the GDF	A	0.06
B1180	so are you	specification	N-A	0.08
B1187	yeah I mean	I suppose	A	0.51
B1198	uh	um so	A	0.89
B1206	uh-huh	to create a new part	A	0.62
B1214	okay	when the parts were created	A	0.67
B1214	so I'm no*	okay	N-A	0.44
B1217	um	um	A	0.92
B1222	yeah but it's just	cause that was a while ago	N-A	0.44
B1238	yeah	a look at it	A	0.53
B1283	right	do it	A	0.11
B1284	yeah yeah	do you know what I mean	N-A	0.82
B1285	so	yeah yeah	A	0.83
B1300	yeah	in the way of things	A	0.01
B1301	one uh once it is	in the way of things	N-A	0
B1309	oh yeah	this way	N-A	0.34
B1327	yep fair enough	I'll look at it	A	0.73
B1328	but I	fair enough	N-A	0.6
B1332	uh what el*	my memory of things	A	0.84
B1338	yeah	properly	A	0.86
B1350	oh yeah uh-huh	to see	A	0.45
B1354	yeah	will be	A	0.69
B1355	so um	will be	N-A	0
B1367	in Dutch	this in	A	0.46
B1367	in Dutch	in Dutch	A	0.89
B1368	yeah	as well	A	0.93
B1386	I may be imagining it	that must have been before me	A	0.78
B1388	well I kno* I know	I may be imagining it	N-A	0.33
B1394	oh so we can	store separately	N-A	0.49
B1395	yeah	the Dutch voice over	A	0.84
B1396	yeah so	the same thing	N-A	0.76
B1403	exactly	out of	A	
B1404	yeah	exactly	A	
B1404	yeah so	out of	N-A	0.91
B1414	yeah	afterwards	A	0.66
B1416	yeah yeah	all at once	A	0.66
B1416	okay the other	all at once	N-A	0.47
B1424	might use	the language that people	A	0.94
B1428	anything with the voice yeah	doing that	A	0.81
B1431	alignment	you start talking about	N-A	0.69
B1437	yeah	or anything like that	A	0.68
B1437	mm-hmm	yeah	A	0.58

B1438	so I think	mm-hmm	N-A	0.7
B1444	boof	boof	A	
B1449	or ding	probably yeah	N-A	0.87
B1450	yes	or ding	A	0.76
B1452	probably actually	I don't know what	N-A	0.63
B1458	yeah yeah yeah	and that does not happen	A	0.55
B1459	mm-hmm-hmm	yeah yeah yeah	A	0.22
B1460	well that	mm-hmm-hmm	N-A	0.02
B1460	so if they make	yeah yeah yeah	N-A	
B1468	yeah	happened yeah	A	0.39
B1473	yeah	there's no voice	A	0.62
B1474	well one thing	yeah	N-A	0
B1484	mm-hmm	or whatever else	A	0.71
B1485	or	mm-hmm	N-A	0.36
B1490	mm-hmm	arrow	A	0.56
B1490	that points	mm-hmm	N-A	0.73
B1499	you can yeah	generate	N-A	0.08
B1502	mm-hmm	whatever	A	0.84
B1505	yeah	big arrows or	A	0.89
B1506	so	or whatever	N-A	0.52
B1509	yes	right	A	0.15
B1510	mm-hmm	when they has thoughts about the	A	0.66
B1511	yeah	language	A	0
B1512	yes	stuff	A	
B1513	but uh	stuff	N-A	
B1518	oh well remind	but	N-A	
B1530	mm-hmm	to it all as well so	A	0.12
B1540	I bet Marloes just talked to 'em	they'd given their	N-A	0.79
B1541	probably	just talked to 'em	A	0.92
B1542	I suspect so	just talked to 'em	N-A	0.6
B1545	ah I think	cause I haven't had a reply back	N-A	0.23
B1546	yeah to	to us to	A	0.55
B1549	is it hard	just in case they had	N-A	0.33
B1555	mm-hmm	the language	A	0.52
B1562	no	something like this	A	0
B1566	mm-hmm	I think	A	0.38
B1582	all right	his Tangram set	A	0.43
B1582	you want to give back	his Tangram set	A	0.31
B1589	yeah it's i* well	helpful to have it	N-A	0.2
B1594	oh	manual	A	0.62
B1594	but it's so	manual	N-A	0.37
B1596	uh yeah	little metal bits	A	0.88
B1597	if you're	uh yeah	N-A	0.01
B1602	mm-hmm	then I'll take it back	A	0.03
B1603	right	then I'll take it back	A	0.84
B1609	yeah	anyway so	A	0.43
B1614	um	plenty of toys	N-A	0.61
B1630	oh similar types	types	A	0.13

B1631	yeah	oh similar types	A	0.29
B1631	so one page	uh-huh	N-A	0.93
B1640	yeah	or style	A	0.73
B1645	right	complexity	A	0.73
B1645	so	complexity	N-A	0.54
B1651	google	google	N-A	0.55
B1651	is there a big	google	N-A	0.7
B1657	mm-hmm	one booklet that came	A	0.82
B1662	or	I'll try	A	0.71
B1663	or is it obvious	you never know what's	N-A	0.54
B1666	well	when you look at these things	N-A	0.59
B1676	yeah	complexity	A	0.52
B1678	but that	for the task	N-A	0.46
B1691	uh-huh	probably three	A	0.68
B1692	uh we've to	uh-huh	N-A	0.76
B1692	so that's not gonna	uh we've to	N-A	0.4
B1700	mm-hmm	you wanna be	A	0.58
B1703	yeah ideally	scale for these things	A	0.92
B1703	if you look at 'em	yeah ideally	N-A	0.01
B1706	mm no	do you just know	N-A	0.82
B1707	no	mm no	A	0.79
B1720	I wouldn't know	people find them or	A	0.68
B1720	right okay	I wouldn't know no	A	0.83
B1729	who in psychology	ranking sites or	N-A	0.64
B1739	I'll try and find out	you can contact	A	0.87
B1742	right	I don't know off	A	0.95
B1742	I only know	I don't know off	N-A	0.7
B1750	um I've not encountered them	do people still use Tangrams	N-A	0.19
B1752	oh so we are very old fashioned	um I've not encountered them	A	0.76
B1761	well	I suppose for	N-A	0.45
B1764	but I'm not sure	right	N-A	0.8
B1766	since then	since then	A	0.75
B1798	yeah yeah	the analysis	A	0.68
B1800	yeah yeah	right	A	0.05
B1814	uh-huh	so what I'm gonna do	A	0.81
B1824	yep mm-hmm	easier to do the	A	0.45
B1824	yeah	the XML	A	0.9
B1831	yeah	the GDF	A	0.71
B1831	cause there are	the GDF	N-A	0.02
B1835	mm-hmm	eye-eye lag	A	0.23
B1860	yeah	sort of setting so	A	0.65
B1860	yeah yeah	yeah	N-A	0.78
B1860	so you might wanna	sort of setting so	N-A	
B1875	yeah	so	A	0.69
B1877	yeah	she'll have views on it	A	0.7
B1877	so but if we could	she'll have views on it	N-A	0.59
B1892	mm-hmm	for it	A	
B1893	um oh I found	for it	N-A	0

B1916	oh yeah	replicate the model out of	A	0.63
B1926	oh yeah who is that then	compared to the hand movements and	N-A	0.42
B1927	Ballard	that then	A	0.93
B1928	kay	Ballard	A	0.92
B1931	are they still on the business	back in ninety two	N-A	0.47
B1933	yes	are they still on the business	A	0.8
B1933	I think so	are they still on the business	N-A	0.66
B1934	where are they	I think so	N-A	0.46
B1935	uh the * I'll *	where are they	N-A	0.11
B1936	not sure	where are they	N-A	0.06
B1937	okay	not sure	A	0.82
B1942	yeah	right	A	0.73
B1946	no	that long ago	A	0.19
B1955	uh that's what I've	a paper you can get	N-A	0.92
B1958	well they still do	and find out	A	0.85
B1960	mm-hmm	n stuff	A	0
B1961	if not	so	N-A	0.13
B1965	uh-huh yep	the National Library is bound to have it	A	0.67
B1977	she's not here next week anyway	for discussion about	A	0.61
B1978	oh well	anyway	N-A	0.71
B1985	okay	before she's back	A	0.77
B1986	oh yeah	before she's back	N-A	0.22
B1992	oh is that what	from her normal account	A	0.48
B1995	yeah	Yahoo account or something	A	0.77
B1996	and I c*	that things are getting forwarded	N-A	0.55
B2010	uh-huh	one message so far	A	0.15
B2021	anything else	anything else	A	0.89
B2039	yeah	anyway	N-A	0.04
B2044	yeah	it's got everything	A	0.54
B2045	mm-hmm	from that	A	0.27
B2046	but I don't think	from that	N-A	0.27
B2052	oh well maybe it is then	I added things to it	A	0.91
B2053	well I don't know	oh well maybe it is then	N-A	0.69
B2059	yeah	two weeks anyway but uh	A	0.63
B2060	uh-huh	two weeks anyway but uh	A	0.59
B2061	I	two weeks anyway but uh	N-A	0.6
B2071	mm-hmm	and JP	A	0.59
B2074	that's what	they're	N-A	0.22
B2081	mm-hmm	want out of that	A	0.04
B2082	yeah	data an	A	0.33
B2088	yeah	he was talking about	A	0.39
B2088	yeah 'cause	so	N-A	0
B2110	yeah	cause a bit of trouble but	A	0.69
B2110	which is yeah	cause a bit of trouble but	N-A	0.82
B2116	oh right	so	N-A	0.07
B2128	yeah	is that	N-A	0
B2137	yeah	and things like that	N-A	0.15
B2139	yeah	through the eye-tracker	N-A	0.01

B2147	yeah yeah	the new parts area	N-A	0.74
B2147	okay	or they are looking at	N-A	0.7
B2149	yeah	that is post analysis	A	0.4
B2150	so right	yeah	A	0.34
B2151	yeah	um	A	0.87
B2151	so the other is	yeah	N-A	0.1
B2163	filter out	from it	A	0.02
B2163	mm-hmm	filter out	A	0.89
B2164	yeah	so	A	0.55
B2204	what documentation	on this project or something	N-A	0.51
B2208	oh okay	JAST report	A	0.36
B2210	well	officially	N-A	0.36
B2214	reality	my salary	A	0.9
B2214	uh-huh	my salary	A	0.75
C1	I mean if you’re ever	yeah it says	N-A	
C4	yeah	it's not	A	0.9
C11	yeah	that	A	0
C11	but I'll	that	N-A	0
C19	really	at the stations	A	0.57
C22	really	passed	N-A	0.04
C22	yeah	really	A	0.71
C23	great	yeah	A	0.37
C26	uh I'm not sure	backup	A	0.09
C32	ah	hill or something	A	0.12
C35	agh	either side	N-A	
C38	yes yes	nasty	A	0.93
C39	yeah	nasty	N-A	0
C44	yeah yeah	so	A	0.37
C71	i don't know what it's doing	or something	A	0.55
C109	yeah	so	A	0.19
C110	I haven't	so	N-A	0
C123	oh yeah uh-huh	generate GDF files	A	0.77
C132	uh-huh	do things	A	0
C149	um	um	N-A	0.02
C159	uh well	a good idea	N-A	0.03
C166	yeah yeah	is standard between	A	0.61
C167	uh well	is standard between	N-A	0
C175	yeah	yeah it'll be alright	N-A	0.85
C175	okay	yeah	A	0.42
C177	i believe	just as	A	0.94
C178	yeah	it's a good strategy	A	0.51
C198	wha	so	N-A	
C216	yeah but I'm	shows you	N-A	0
C225	okay	DTD	A	0.54
C226	ju* ju*	yeah	N-A	0.16
C231	okay	format	N-A	0.08
C234	yeah	tagnames later	A	0.43
C235	okay	yes	A	0.42

C245	yeah	in the end	A	0
C245	yeah	in the end	A	0.91
C245	no I'm just	in the end	N-A	0
C263	um	or	N-A	0
C280	uh-huh	the moment	A	0.23
C284	yeah	format	A	0.46
C300	yeah	right	A	0.32
C329	yeah	point	A	0.77
C330	so are	well	N-A	0.4
C359	oh okay	generate	N-A	0
C366	uh	to do	N-A	0.04
C366	not yet	no	N-A	0.89
C367	okay	no	A	0.45
C367	but out a	okay	N-A	0.62
C373	yeah	for it	A	0.26
C382	uh-huh	whatever	A	0
C388	okay	GDF	A	0.77
C389	so this	stuff	N-A	0
C397	no	right	A	0.05
C399	but ther was a number	don't think so	N-A	0.44
C411	yeah	the pipeline order	A	0.07
C415	yeah	you mentioned	A	0.01
C415	um	you mentioned	N-A	0.38
C416	and I'm not sure	um	N-A	0.42
C420	yeah	things to do	A	0.76
C421	um	things to do	N-A	0.26
C434	until you get the	working	N-A	0
C434	microphones	until you get the	A	0.59
C435	oh microphones	microphones	N-A	0.24
C439	batteries	we're waiting for	N-A	0.42
C443	okay	beginning this week	N-A	0.24
C448	have you tried maplin	quite strange	N-A	0.47
C454	are they expensive	has ordered them from	N-A	0.22
C459	cause you know	or something like that	N-A	0.19
C465	uh	getting in the way	N-A	0.59
C474	so	today	N-A	0.05
C489	right	correctly	A	0.04
C490	it's just	to check	N-A	0
C493	not yet	this is	N-A	0.12
C494	yeah	not yet	A	0.68
C521	oh yeah	so	N-A	0.01
C529	yeah	down there	A	0
C530	I acrually	yeah	N-A	0.59
C546	yeah	Haymarket	A	0.6
C556	ah yeah	I know where it is off	A	0.81
C563	yeah	where maplin is	A	0.88
C563	I just think	where maplin is	N-A	0.8
C570	no it's	a hundred	A	0.42

C572	yeah	I need	A	0.97
C580	uh no	recording	A	0.59
C583	yes	know this	A	0.6
C583	I suppose	yes	N-A	0.01
C596	right yeah	funny down there	A	0.48
C601	cause you know	call 'em	N-A	0.38
C613	so	isn't it	N-A	0
C645	can't remember them	before you could run	N-A	0.71
C664	uh his	and	N-A	0.05
C668	yeah	october	N-A	0.84
C676	ah yeah	october f*	N-A	0.65
C681	I think he's	he's behind hand	A	0.83
C696	oh well they're probably away then	dead for almost last two weeks	A	0.49
C701	no	at the same place	N-A	0.01
C706	oh	review	A	0.46
C713	oh you have	for	N-A	0.03
C721	yes	right	A	0.77
C724	well we don't need to	and then you	N-A	0.46
C729	okay	anyway	A	0.55
C738	yep	type	A	0.96
C740	okay	task	A	0.27
C740	um	task	N-A	0.21
C761	okay I think	anyway	N-A	0.55
C766	that's newer then	remember exactly what it was	A	0.77
C795	yeah yeah	from the model	A	0.14
C804	okay	or from joe	A	0.77
C805	so I've got	or from joe	N-A	0.43
C815	yep	done done done	A	
C815	anyway	yep	N-A	
C821	yeah	oh is he back	A	0.75
C821	ah	yeah	A	0.93
C831	well he's just	for him at the moment	N-A	0.61
C870	yeah I mean	was a mirror image	A	0.74
C870	I'm considering	was a mirror image	N-A	0
C882	well I	the old one so	N-A	0.09
C920	yes	for your experiment	A	0.25
C923	yes	from marloes	A	0.77
C923	or at least	from marloes	N-A	0.43
C930	aha yeah yeah	have actually run	A	0.81
C937	right	all very quiet	A	0.74
C938	have you tried	all very quiet	N-A	0.82
C943	uh	I need things that	N-A	0.84
C950	ach try phone	type of uh	N-A	0.4
C966	oh right	today I think	N-A	0.15
C968	um	be ready by then	N-A	0.14
C974	right	hoping	A	0.47
C975	okay	hoping	A	0.47
C975	we have	okay	N-A	0.18

C980	uh-huh	down	A	0.07
C981	yeah	down	A	0.63
C992	yeah	december I think	A	0.54
C992	so	december I think	N-A	0.5
C997	um	getting them	N-A	0.24
C1011	yeah	users	A	0.85
C1017	by	again	N-A	0
C1018	Jules	by	A	0.23
C1047	lecturer	is it a	A	0.65
C1049	right	lecturer	A	0.71
C1054	well	the entire	N-A	0.28
C1057	block book for how long	sort of	N-A	0.25
C1065	uh-huh	six days	A	0.52
C1069	uh-huh	of those days	A	0.13
C1100	mm-hmm	time	A	0.12
C1100	yeah	time	A	0.77
C1116	yeah	booking system	A	0.26
C1116	yeah	yeah	A	0.76
C1123	mm-hmm	meeting	A	0.76
C1132	in theory	use of the lab	N-A	0.2
C1135	oh really	every morning	N-A	0.02
C1155	no	connections or	N-A	0
C1157	no	or anything	A	0.67
C1158	okay	no	A	0.4
C1174	yeah	should	A	0.52
C1193	oh well	the lab	A	0.64
C1198	uh-huh	is there	A	0.1
C1201	mm-hmm	yeah	A	0.89
C1209	yeah	you would anyway	A	
C1209	so	yeah	A	
C1224	should be	you're happy	A	0.55
C1226	oh	got to get	N-A	0.82
C1238	oh yeah	video	N-A	0.26
C1239	uh-huh	video	A	0.27
C1256	yeah but you	so	N-A	0.01
C1264	yeah	trial	A	0.05
C1264	yeah okay	yeah	N-A	0.43
C1291	yeah	when I can	A	0.69
C1291	hopefully	when I can	N-A	0.62
C1308	it hasn't been	gonna be	N-A	0.76
C1310	well	in that sort of	A	0.43
C1310	what I'm trying to get at is	in that sort of	N-A	0.65
C1325	uh-huh	camtasia	N-A	0.04
C1325	yeah	camtasia	A	0.76
C1338	it shouldn't	not that hard I mean	N-A	0.35
C1347	yeah	complexity	A	0.88
C1377	yeah	to do	A	0.91
C1380	oh yeah	hoping	N-A	0.82

C1384	you can	exactly	N-A	0
C1406	yeah	unless you're	A	0
C1411	but presumably	today	N-A	0.48
C1414	I think	over	N-A	0.06
C1418	mm-hmm	presumably	A	0.83
C1419	yeah	free	A	0.65
C1421	I d*	or	N-A	0
C1428	um	that's all	N-A	0.27
C1429	well	that's all	N-A	0.62
C1436	yeah	models	A	0.57
C1440	that's quite	optimal	N-A	0
C1448	mm-hmm	constructing	A	0.87
C1452	yep	which	A	
C1463	um	experiment schedule so	N-A	0
C1471	nn-hnn	for that	A	0.11
C1476	you gonna run out of	um	N-A	0.33
C1482	yeah	might take a wee while	N-A	0.46
C1491	yeah	part of it	A	
C1496	yeah	the pair of you	A	0.76
C1497	yeah I mean	the pair of you	A	0.5
C1497	I think	the pair of you	N-A	0.01
C1506	yeah	easiest	A	0.71
C1508	it's just what I'm	um	N-A	0
C1525	uh-huh	book in an	A	0.02
C1533	yeah	in the lab	A	0.79
C1534	cause he might	uh	N-A	0
C1536	oh yeah	let them get	N-A	0
C1536	no no no	started	N-A	0.05
C1567	mm-hmm	outputter	A	0.22
C1567	yeah	outputter	A	0.84
C1571	yeah	as well	A	0.74
C1586	what I do*	parts	N-A	0
C1589	yeah	part builder	A	0.47
C1591	yeah	the investment	A	0.47
C1592	yeah	right	A	0.67
C1592	well I'll have	right	N-A	0
C1602	yeah yeah yeah yeah	symmetrie	A	0.82
C1604	for	before	A	0.8
C1605	so if I	for	N-A	0.01
C1610	yeah but not	a few bugs	N-A	0
C1616	no	doing it	A	0.63
C1617	no	no	A	0.83
C1620	yeah	so	A	0.6
C1621	well	um	N-A	0
C1630	yeah	part builder	A	0.69
C1630	yeah	part builder	A	0.72
C1631	but I'm not	part builder	N-A	0
C1642	right	certainly	A	0.63

C1646	yeah	on it	A	0.46
C1647	yeah	worth it	A	0.95
C1648	you	yeah	N-A	0
C1659	oh really	task list	N-A	0
C1664	right	else	A	0.64
C1674	right	this NXT stuff	A	0.48
C1675	yeah well that	so	N-A	0.19
C1679	oh yeah yeah yeah	GDF	A	0.64
C1683	yeah yeah	you know	A	0.94
C1695	right	works	A	0.8
C1697	cause that's	you know	N-A	0.41
C1697	there's stuff to learn there	cause that's	N-A	0.2
C1705	yeah yeah	it fits what	A	0.69
C1707	yep	so	A	
C1714	yeah	I'm sure	A	0.6
C1715	so just don't	I'm sure	N-A	0.25
C1722	yeah	to come	A	0.66
C1722	we'll find	yeah	N-A	0.19
C1742	okay	GDF	A	0.74
C1743	so I	GDF	N-A	0.62
C1747	yep	get together	A	
C1760	yeah	next stage	A	0.57
C1777	oh	jules	N-A	0
C1788	oh okay	next wednesday	A	0.22
C1801	yep	twos	A	0.2
C1811	oh yeah	teaching	A	0.04
C1812	that should be fun	teaching	N-A	0.38
C1844	oh yeah	so	N-A	0
C1861	yeah	it's just	A	0.72
C1870	yeah	everywhere	N-A	0
C1874	yeah	by doing this	A	0.61
C1874	well	by doing this	N-A	0.75
C1882	I'm pretty sure	pulled it so far	N-A	0.62
C1884	well I	in my day so	N-A	0.08
C1887	really	so	N-A	0.35
C1889	in april	started here	A	0.58
C1894	yeah I mean	I'm sure	N-A	0
C1894	I got	yeah I mean	N-A	0.9
C1904	oh well	but that's all	A	0.75
C1904	yeah I can remember	oh well	N-A	0.27
C1919	oh yeah ah	take a charge of	A	0.39
C1938	okay	queation	A	0.02
C1957	yeah	or something	A	0.65
C1970	yeah	different tiers	A	0.08
C1973	yes	as well	A	0.25
C1975	yeah this	the eye	N-A	0
C1988	cause a lot	yeah I guess it's	N-A	0.24
C1991	as well	language part	A	0.83

C1993	yeah	yeah	A	0
C1993	so I was just	yeah	N-A	0.94
C1998	yeah	would be the	A	0
C2017	um	right	N-A	0.66
C2043	well	the fact	N-A	0.03
C2047	mm-hmm	track	A	0.94
C2048	and uh	mm-hmm	A	0.62
C2060	mm-hmm	right	A	0.44
C2071	yeah	gaze	A	0.57
C2071	but	gaze	N-A	0.07
C2079	so it may	the way they're looking	N-A	0.04
C2085	mm-hmm	account	A	0.48
C2107	mm-hmm	know about	A	0.4
C2127	yes	thing	A	0.53
C2137	oh right okay	previous fixation	A	0.3
C2140	cause you are not	in that case	N-A	0.88
C2146	right	deciding factor is	A	0.76
C2147	uh-huh	deciding factor is	A	0.79
C2148	in that case	deciding factor is	N-A	0.18
C2154	yes	corrected	A	0.35
C2154	and then	yes	N-A	0.57
C2164	yeah	just a mark	A	0.2
C2165	yeah	position	A	0.43
C2165	so	position	N-A	0.61
C2189	hmm	set of codes	A	0.56
C2189	no	set of codes	N-A	0.45
C2192	ah yeah	no	N-A	
C2203	uh I think	um	N-A	0.21
C2224	right	what	A	0.82
C2234	right	stream	A	0.9
C2240	right okay	as a	A	0.55
C2241	well i*	as a	N-A	0.59
C2273	right	time out	A	0.76
C2273	yeah	time out	A	0.63
C2274	but what we need	time out	N-A	0.34
C2282	yeah	automatically	A	0.09
C2296	yeah	different way	A	0.44
C2296	yeah 'cause I think	different way	N-A	0.31
C2300	yes	information	A	0.82
C2304	yeah ther's a blink	documantation I've read through	A	0.65
C2305	yeah	yeah ther's a blink	A	0.78
C2306	but it could be	sort of	N-A	0.09
C2313	possibly	so	A	0.72
C2344	it shoud just	big problem	N-A	0.09
C2349	yeah	or something anyway	A	0.37
C2353	if they're not	so	N-A	0
C2371	oh right yeah you said that	absorbing the new information	A	0.71
C2384	well	so	N-A	0.15

C2391	uh it	saccade	N-A	0
C2406	right	output	A	0.71
C2417	okay	that sort of thing	A	0.3
C2419	I kinda get this	is a	N-A	0.55
C2422	right	viewing time	A	0.88
C2432	and so is there	okay	N-A	0.8
C2438	yep	anything	A	0.31
C2438	although	anything	N-A	0.01
C2453	yeah	or not	A	0.62
C2453	cause I think	or not	N-A	0.17
C2460	yeah	to NXT	A	0.01
C2505	right	analyses	A	0.14
C2519	well that's	analyses	N-A	0.33
C2550	yeah	linguistic	A	0.63
C2554	right	the actions	A	0.85
C2574	yeah	can't you	A	0.76
C2574	I mean	can't you	N-A	0
C2615	right	does	A	0.25
C2616	okay	does	A	0.8
C2616	okay	does	A	0.63
C2616	cause you	okay	N-A	0.37
C2631	right	so	A	0.85
C2654	oh yeah	to lump	N-A	0.07
C2682	yep	on that	A	
C2684	so	fair enough	N-A	0.45
C2701	yes	starting point	A	0.62
C2724	fine	um	A	0
C2724	good	um	A	0
C2735	cause it will	that works	N-A	0.12
C2749	yeah	junk	A	0.64
C2749	ah	junk	A	0.51
C2755	we don't wanna	um	N-A	0.03
C2761	fine	correct	A	0.09
C2763	yeah	possibly	A	0.12
C2763	well if is	possibly	N-A	0
C2771	yeah	doing the GDF	A	0.76
C2772	I'm just	so	N-A	0
C2777	yeah	ASCII	A	0.06
C2801	ah yeah	break our	A	0.42
C2824	yeah yeah but	that should be interesting	A	0.75
C2829	uh-huh	robin	A	0.25
C2837	right	java	A	0.58
C2855	right	so	A	0
C2865	yeah	or	A	0.68
C2866	fair enough	or	A	0.72
C2866	wel so s*	or	N-A	0.18
C2870	oh yeah	wouldn't it	A	0.49
C2870	and it's easy enough	wouldn't it	N-A	0.48

C2873	yeah	track	A	0.9
C2874	we gotta	yeah	N-A	0.32
C2881	okay	week	A	0.5
C2882	johnathan	okay	N-A	0.56
C2898	yeah so	going through	A	0.81
C2902	yeah	so	A	0.93
C2903	he is a cool	looking at that	N-A	0
C2911	yeah	right	A	0
C2923	sorry	blank still	N-A	0.21
D48	yeah	at some finer	A	0.76
D53	yeah	and a method of	A	0.85
D82	yeah	because you treated it so grossly	A	0.48
D192	okay	at some time slice rate	A	0.71
D207	mm-hmm	over time	A	0.31
D234	mm-hmm	right	A	0.46
D248	yeah	it's gonna be on	A	0.23
D263	mm-hmm	I stayed here	A	0.49
D319	mm-hmm	in which two people are looking at the same thing	A	0.85
D330	mm-hmm	one could imagine those two categories	A	0.83
D355	mm-hmm	alright	A	0
D380	yeah	and their mouse movements and things	A	0.49
D391	yeah	basically get	A	0.57
D394	okay	from that	N-A	0.81
D445	mm-hmm	typical construction event	A	0.83
D459	mm-hmm	a construction already begun	A	0.7
D468	mm-hmm	right	A	0.74
D480	mm-hmm	where they're looking	A	0.06
D489	mm-hmm	which could be zero	A	0.48
D495	mm-hmm	which is next going to be added	A	0
D547	right	our measure of alignment	A	0.86
D564	so	so	A	0.65
D686	yep	if your	A	0.81
D717	yeah	yeah sure	A	0.04
D722	mm-hmm	at these	A	0
D796	yeah they're	that one is looking	A	0.78
D797	yep	the other is looking the same	A	0.41
D805	mm-hmm	that's true	A	0.79
D816	yeah	at the clock	A	0.56
D821	so	sure it's a seperate	N-A	0.31
D830	yeah	where the eye is	A	0.93
D835	mm-hmm	any interesting category	A	0.79
D842	yeah	it's um:	A	0.21
D874	but the ex*	so	N-A	0.1
D930	but we believe	he's actually	A	0.89
D956	yeah	and we should just apply it	A	0.36
D960	mm-hmm	they do a lot of sub-assembly	A	0
D976	I'll just I'll just minute that	I don't know	N-A	0.58

D1114	it's a possibility	but this is but	A	0.56
D1179	really	give you everything you need to know	N-A	0.31
D1182	okay	that's the claim	A	0.58
D1184	anyway	right	N-A	0.01
D1334	what does that mean	by the only way we're gonna use frame rate	N-A	0.08
D1385	okay	steady fixation	A	0.78
D1627	yeah	presumably	A	0.8
D1734	mm-hmm	in the triangle region	A	0.75
D1766	yeah	maybe a mouse pointer	A	0.84
D1773	I don't know	in the same place	N-A	0.77
D1823	yeah	when the mouse isn't also there	A	0.52
D1843	mm-hmm	to the construct	A	0.8
D1852	mm-hmm	right	A	0.4
D1867	mm-hmm	B's gaze	A	0
D1869	yeah	likewise	A	0.75
D1935	yeah	where the mouse is	A	0.59
D1985	does that happen	about that	N-A	0.15
D1989	oh right	fuzz factor	A	0.87
D1993	uh-huh	close together	A	0.01
D2057	what do you mean by greatest overlap	what they are actually looking at	N-A	0.48
D2102	uh-huh	there	A	0.77
D2110	yeah	the two pieces simultaneously	A	0.69
D2148	right	it'll be an actual	A	0.84
D2159	mm-hmm	target	A	0.7
D2293	well we'll see if ELAN likes it or not	what's going on	A	0.72
D2436	yeah	when you create a new part	A	0.81
D2526	mm-hmm	to the triangle	A	0
D2592	mm-hmm	right	A	0.7
D2735	nn-hnn	it's a beatifully controlled situation	A	0.93
D2741	mm-hmm	if everything is purple	A	0.52
D2775	yeah	mold	A	0.86
D2796	right okay	that specifies the parts	A	0.5
D2827	mm-hmm	in the configuration	A	0.3
D2846	has multiples yeah	and Boufix	A	0.78
D2855	mm-hmm mm	and all the	A	0.06
D2883	yeah	and I think	A	0.18
D2885	mm-hmm	that's how the robot is set up to	A	0.06
D2919	yeah	at triangle one	A	0.87
D2920	so	when is triangle one	N-A	0.15
D2948	what do you mean define	that people build	N-A	0.37
D2972	mm-hmm	I mean	A	0.97
D3000	mm	in the history of construction	A	0.6
D3126	mm-hmm	and when they don't	A	0.42
D3153	oh yeah you said that	so you can't collect them	A	0.51
D3177	uh-huh	sort of thing	A	0.26
D3183	mm-hmm	they have to bting it into play	A	0.87
D3192	yeah where there is extra parts	or anything	A	0.74

D3204	yeah	so you can't have a	A	0.88
D3225	yeah	so	A	0.74
D3263	but that's a human coding	one of them says somethin	N-A	0.75
D3269	mm-hmm	just from the	A	0.6
D3296	yeah	they say	A	0.02
D3345	mm-hmm	we'll get a bad score	A	0.78
D3435	uh-huh	it seemed to interfear with it	A	0.53
D3708	right okay	what I was assuming and and	A	0.69
D3770	uh-huh	to merge into a stereo	A	0.81
D4079	mm-hmm	is sometimes a good idea	A	0.62
D4129	yeah	in an anechoic room	A	0.74
D4227	yeah yeah	you just have to write read	A	0.76
D4807	mm-hmm	for next week	A	0.76
D4859	okay	that somewhat	A	0.23
D4896	mm-hmm	alright	A	0.8