

Pitch Contour Matching and Interactional Alignment across Turns: An Acoustic Investigation

Jan Gorisch, Bill Wells and Guy J. Brown

University of Sheffield, UK

Abstract

In order to explore the influence of context on the phonetic design of talk-in-interaction, we investigated the pitch characteristics of short turns (insertions) that are produced by one speaker between turns from another speaker. We investigated the hypothesis that the speaker of the insertion designs her turn as a pitch match to the prior turn in order to align with the previous speaker's agenda, whereas non-matching displays that the speaker of the insertion is non-aligning, for example to initiate a new action. Data were taken from the AMI meeting corpus, focusing on the spontaneous talk of first-language English participants. Using sequential analysis, 177 insertions were classified as either aligning or non-aligning in accordance with definitions of these terms in the Conversation Analysis literature. The degree of similarity between the pitch contour of the insertion and that of the prior speaker's turn was measured, using a new technique that integrates normalized F0 and intensity information. The results showed that aligning insertions were significantly more similar to the immediately preceding turn, in terms of pitch contour, than were non-aligning insertions. This supports the view that choice of pitch contour is managed locally, rather than by reference to an intonational lexicon.

Keywords

conversational alignment, pitch contour, pitch matching, prosodic repetition, prosodic similarity

Introduction

There has been relatively little engagement with pitch contrastivity and its possible functions within 'phonetics of conversation' research. This is somewhat surprising given that, in conversation analysis (CA) more widely there are widespread assumptions about the meaning of pitch/tone direction evidenced by their embodiment in the Jefferson transcription system (Jefferson, 2004) and the use in passing of phrases such as "try marked intonation" (Schegloff, 1996, p. 101). In one of the rather few published analyses of the contrasting meanings of different tones, Gardner (1997, 2001) found

Corresponding author:

Jan Gorisch, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK

Email: J.Gorisch@sheffield.ac.uk

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 an *mm* allows the previous speaker to continue talking (cf. Schegloff, 1982). 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).

However, at least as frequent as such descriptions of consistent mapping from pitch contour to conversational function are reports of the high degree of variability in the pitch contours used by speakers when producing the same social actions in spontaneous conversation. Geluykens (1987) illustrates the unpredictability of tone direction in relation to pragmatic distinction between questions and statements in English. Szczepk Reed (2004) has suggested that pitch contour may vary in rather unsystematic ways, in relation to pragmatic function, at the end of turns in naturally occurring talk-in-interaction. Kaimaki (2011) reports apparently free variability between rises and falls in initial turns in phone conversations. This kind of variability is also demonstrated in an investigation of the potential opposition between pitch contours at the same place in interactional structure that was carried out by Walker (2004). Walker examined the phonetic characteristics of adjacency pairs in a corpus of naturally occurring conversational data. The first pair parts included invitations, enquiries, offers, assessments and requests. Syntactically, both interrogative and declarative forms were found and two distinct pitch contours were also found, one falling and the other rising. However, there was no evidence of any relationship between the syntactic form of the first pair part and its pitch contour; nor between pitch contour and the type of first pair part, for example whether it was a request as opposed to an assessment. Thus neither syntactic nor pragmatic accounts of the meaning of English tones were supported. In sum, the accumulated evidence from analysis of corpora of naturally occurring talk does little to support Levelt's (1989) claim, that tone is used to convey the illocutionary force of an utterance – at least if we understand that to mean that speech acts such as 'request', 'offer' and 'assessment' are predictably related to tonal choices from the intonation system such as 'rise', 'fall' and 'fall-rise'.

However, this does not mean that tone is therefore interactionally irrelevant. An alternative hypothesis is that a speaker's choice of tone is explicable not by reference to an intonational lexicon (in which rises have distinct meanings from falls, for example) but instead by reference to the tone used by the speaker of the previous turn. This is supported by a growing body of research demonstrating the interactional relevance of prosodic orientation by a next speaker. 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. Couper-Kuhlen (1996) demonstrated that F0 contour matches with relative vs. absolute F0 register perform different conversational actions: matching relatively (i.e., with respect to the individual's voice range) contextualizes verbal repetitions as quotation, whereas matching absolutely the F0 of the prior speaker contextualizes the repetition as mimicry. In an analysis of continuers in Italian conversation, Müller (1996) found that two different 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" (p. 163).

In the most wide-ranging study of this phenomenon to date, Szczepk 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 Szczepk 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, news receipts, ‘oh’ and related exclamations and disagreements. More recently, Szczepk 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. (Szczepk 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, 2010, p. 261).

1.1 Phonetic issues

Studies of prosodic matching necessarily rely on some kind of phonetic analysis as the basis for their claims. However, the phonetic identification of a prosodic ‘match’ is a far from trivial matter: questions arise as to what phonetic features or parameters should be regarded as relevant to matching; and how to deal with the obvious individual differences between speakers who nevertheless may be heard to be matching one another. Szczepk Reed (2006) observed prosodic matching in relation to “intonation contour, pitch register, pitch step-ups, loudness, speech rate, voice quality and sound production” (p. 35). Sometimes these observations are supported by analyses based on the display of F0 contours that are logarithmically scaled but are not normalized to take account of differences between speakers.

In Couper-Kuhlen (1996), two visual F0 representations are used to illustrate the acoustic analysis. Because the aim of her study is to investigate absolute vs. relative pitch register matching, the display alters the representation of F0 according to (a) the common base for both speakers with all Hz values expressed as semitone intervals from 50 Hz and (b) the base for the individual speaker’s voice range expressed in semitone intervals from the lowest Hz value that a given speaker “is inclined to use” (Couper-Kuhlen, 1996, p. 374). Depending on which scale captures the match between the two speakers (in her examples male and female), Couper-Kuhlen identifies it as an *absolute pitch register match* (on common base) or a *relative pitch register match* (on individual base). The identification of the latter depends on having a strategy for speaker normalization.

The issue of normalization has been addressed in more recent research into naturally occurring talk. For instance, Heldner, Edlund and Hirschberg (2010) normalized for individual differences and gender in an investigation of speaker transitions. Comparison of backchannels, smooth switches and pause interruptions, with speech following the backchannel showed that the backchannels themselves were most similar in F0 height to the first speaker’s preceding talk. Heldner et al. thus showed that speakers may match each other in F0 height for interactional purposes. However, their averaging method disregards any F0 *movements*, and thus potentially ignores pitch contour matching of the kind that was identified by Couper-Kuhlen (1996), Müller (1996) and Szczepk Reed (2006, 2009).

1.2 Interactional issues

Just as the phonetic analysis of matching presents important challenges, there are issues in identifying the interactional work that matching may be implicated in. As described above, researchers

have proposed that prosodic matching is used for continuing the project in hand, aligning or affiliating with the previous speaker and/or the previous speaker's agenda; whereas non-matching would indicate initiating a new project, disaligning or disaffiliating from the prior speaker and/or his agenda. Until recently, these terms have been used somewhat imprecisely, even interchangeably. However, recently, Barth-Weingarten (2011) has differentiated the terms (dis)alignment and (dis)affiliation as follows: *(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 derives 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' behavior 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" (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" (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 are close to Stivers' and Barth-Weingarten's conceptualization 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 that includes cases of disalignment or misalignment as described in earlier research.

1.3 Motivation for the current study

In summary, several studies have suggested that prosodic orientation to the prior speaker provides the current speaker with a resource for accomplishing social actions; and one important type of prosodic orientation is prosodic matching. If this is indeed the case, then it has potentially far-reaching implications for our understanding of how intonation works. It suggests that the speaker-based, largely context-free models of intonation production that have dominated recent theorizing are at best only partially true. Instead, it will be necessary to take account of the likelihood that the production of a particular intonation pattern is context dependent, being conditioned by the intonation pattern of the immediately prior talk. This in turn has implications for how children learn intonation, and indeed, what it is that they learn (Wells, 2010). However, for this view to be taken

seriously, it is important to demonstrate first that prosodic matching can be robustly and objectively identified; and second that matching and non-matching are devices that are systematically used by participants for interactional ends.

In the majority of the phonetic studies reported above, acoustic analysis is primarily used as objective evidence to support the analyst's hearing that the second speaker's turn matches that of the prior speaker. This rests on some generally accepted assumptions, for example that there is a correlation between measured F0 and perceived pitch; and similarly between measured intensity and perceived loudness. While there will be cases where the hearing of a prosodic match and the matching of acoustic records are evidently mutually supportive, there remains a tricky conceptual problem, articulated by Couper-Kuhlen (1996, p. 368), of what counts as a prosodic match "for all practical purposes". There are at least two aspects to this problem. First, there may be cases where participants appear to treat a turn as a prosodic match in terms of the action it performs, and it may be hearable to the analyst as a prosodic match; but it is hard to identify it as a match from the acoustic records (i.e., 'false negatives'). Second, there may be cases where the acoustic records indicate that the second turn is a prosodic match of the first, but it is not treated as such by the participants in the talk, and may not even be heard as such by the analyst ('false positives').

Testing the robustness of the concept of prosodic matching involves identifying the interactional work that is done by prosodic matching vs. non-matching, and then developing an objective means for distinguishing matches from non-matches. Here, we specifically address the question of how two adjacent pitch contours, produced by different speakers in naturally occurring conversation, can be identified as being matches. Our broader aim is to arrive at an interactionally grounded account of prosodic matching that is supported by objective acoustic analysis.

In line with earlier studies of the phonetics of talk-in-interaction, CA is used in conjunction with detailed phonetic analysis (cf. Local & Walker, 2005). Our focus is on short utterances produced by a second speaker, either in the clear or in overlap, which are preceded by a more extended turn from a first speaker and immediately followed by another turn from that first speaker. These will be referred to as *insertions*. The research is driven by the following question: *is pitch contour matching of an insertion and the immediately preceding turn used for alignment, and is non-matching used for non-alignment?*

2 Material

Insertions were collected from three meetings of the AMI meeting corpus (<http://corpus.amiproject.org/>). This corpus consists of round-table meetings recorded on individual headset microphones and individual video cameras. The corpus 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.

Several methodological factors constrained 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 that involved a consistent set of speakers.
- b) Meetings were chosen in which the speakers 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 that we selected are designated EN2009b (51 minutes in length), 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 accent; speaker B is a female data processing specialist with an American accent who had been living in Edinburgh since 1988 (the recording was made in 2005); speaker C is a male postdoctoral psychologist with a Scottish accent; and speaker D, who is only present in meeting EN2009d, is a female senior psychologist with an American accent. Their corresponding identification numbers in the AMI meeting corpus are MEE094 (A), FEE083 (B), MEE095 (C) and FEE096 (D). These 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, for example as a progress report, to which co-participants may respond. This reporting and discussing environment is comparable to the storytelling situation analyzed by Stivers (2008) in some respects, for example in the asymmetry of contributions from participants.

3 Method

We have restricted our analysis to insertions produced by speaker B. While our eventual aim is to identify practices that are common across speakers, at this stage it facilitates the acoustic analysis to focus on candidate prosodic matches produced by a single speaker. In total, 280 insertions produced by speaker B were collected from the three meetings, of which 177 could be used for the acoustic analysis. The collection of insertions includes tokens of standalone “uh huh”, “oh yeah uh huh”, “uh huh yeah”, “right uh huh”, “right okay”, “oh right yeah you said that”, “oh really” as well as others such as “by” and “until you get the”. (See supplementary document for a complete list: <http://las.sagepub.com/content/55/1/57/suppl/DC1>).

Each insertion, together with its surrounding context, was analyzed by reference to the original recording and the transcript. 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 (see 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 vocalizations 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. A comparison of the orthography and the phonetic characteristics of individual tokens indicates that *uh huhs* have 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 vocalization. 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 vocalization is perceived as having a non-modal voice quality in the middle: breathy [əəə], creaky [əəə], or with aspiration [əʰə]. This

splits the vocalization into two parts that can be described as syllables. The end can also appear to have aspiration [ə^bə^b]. Glottal stops are not generally observed at either the onset or the offset of the *uh huh*. Different vowel qualities can also be found, for example more open [e^be] or more fronted [ɛ^bɛ]. The token *uh huh* can be distinguished from other vocalizations that are based on a schwa-like quality, for example hesitation *uh*, by its bisyllabic structure. Other tokens with two syllables may be nasal throughout: bilabial: *mm hm* [m^bm] or alveolar: *nn hn* [n^bn].

3.1 Interactional analysis

The transcript of each insertion token was examined together with its context, but without reference to its prosodic features, in order to determine whether or not speaker B aligns with the current speaker's agenda in progress. We now illustrate this by reference to four extracts that contain insertions. The identifiers for the extracts contain the letter B, C or D for the meetings (EN2009b, EN2009c, EN2009d) and the time index (in seconds) at which the insertion occurs.

Extract (1) is an instance of an aligning insertion: "uh huh" is used here as a simple continuer. For this extract, following Goodwin (1980), above each line containing the vocal aspects of spoken language, up to two more lines contain a gesture/gaze layer: one for the current speaker and another for the principal recipient. They contain manual, head and other body gestures and the direction of gaze, with commas indicating when gaze shifts. This illustrates the richness of nonverbal activity in the recordings, reflected in our working transcripts, and its potential relevance in making the case that an insertion is aligning or not. In the interests of simplicity and space this level of detail is omitted from the remaining extracts presented here.

Extract 1 (C280)

gaze down-----,,--gaze to C-----((gesticulation with right hand directed to C))----
-----((C: nod, blink))-----

1 A: u:m in the output format for the for the task you know so
-----,,---gaze down-----,,---mid gaze---,,-----
-----((C: nod))-----

2 what what you're what you're gonna do the analysis on .hhh um so
-----gaze to C-----,,--gaze down-----,,--gaze to B--((gesticulation))-----

3 making sure that .h I can take (0.2) get that information out of
-----((gesticulation))-----((gesticulation hold for a moment))

4 the GDF as it st* as a state of the moment .hhh
(((nod, blink)))

5 B: [uh huh
-----((gesticulation))-----

6 A: [and if I find something that you can't get out of it then I can
---((end gesticulation))-----

7 add that to the add that to the GDF format (0.7)

Speaker A in lines 1 to 4 is making an extended report. In line 4, speaker A is still gesticulating when approaching a transition relevance place (TRP). When he reaches it, gesticulation comes to a pause although the hand is still in the air (gesture hold). An audible inbreath follows. The insertion *uh huh*, accompanied by a nod and a blink, from speaker B in line 5 does not accomplish any

other work than handing back the floor to speaker A, who in line 6 comes back in overlap to continue his talk on his agenda and gesturing. Thus the *uh huh* is treated by B and A as a continuer: it aligns with the action in progress, which is A's reporting.

Another insertion by B that aligns with speaker A can be seen in Extract (2). Alignment is first indicated by the lexical content of the insertion "ah yeah", which routinely (though not inevitably) expresses agreement with the prior speaker's talk; and second by speaker A's treatment of the insertion to allow him to continue on his agenda: line 4 is not addressing the insertion in any other specific way.

Extract 2 (C556)

```

1   A: ... to Maplin I kno* I know where it is off
2       I [used to live there
3   B: [ah yeah
4   A: so (0.7) I know [I know where .hh
5   B: [oh yeah
6       yeah

```

On the other hand, Extracts (3) and (4) exemplify non-aligning insertions.

Extract 3 (C1017)

```

1   C: ... u:hm the lab has been block booked again
2   B: by::
3   C: Jules (0.5) uh from: u::h psychology hh
4   B: um (0.7) I don't know this person (0.5) so (1.0) they (1.0) well ...

```

In (3), speaker C reports that a laboratory "has been block booked again", which prevents the researchers using it for their purposes. Following the insertion "by:", C continues to talk, but his turn "Jules" is only understandable as a second pair part responding to "by" as a question (i.e., an "increment initiator" (Lerner, 2004)). Thus B's insertion serves to non-align with speaker C's own construction of the progression of his report, requiring C to expand on his prior talk in a quite specific way.

Extract 4 (C1135)

```

1   B: ... she and Martin may have had discussions a [:bout
2   C: [.h uh
3   B: appropriate use of the la:b h
4   C: uh in theory JAST is meant to have it every morning hh
5   B: (0.5) oh really [(smile))
6   C: [hh ((nod))
7   B: yeah but I g* don't you guys set up for (.)
8       isn't block booking more appropriate (.) becaus::e (1.5)
9       i* isn't ...

```

In (4), speaker B is concerned by the lab booking schedule, which seems to be inappropriate. Her statement expressing her concerns (lines 1 to 3) is followed by a statement by C that endorses these concerns, as the project is "meant to have it every morning". B's insertion "oh really" in line 5

serves as a first pair part, as shown by the fact that C's resumption takes the form of an affirmative nod (line 6). Thus in this instance, B's insertion does not align with C's reporting; she subsequently challenges his statement of having the lab "in theory" as not "appropriate" enough (lines 7, 8). Short insertions such as *oh really* can be used to express surprise and may initiate repair from the prior speaker. As Selting (1996) shows, such astonished repair initiators then require special treatment by the prior speaker. In the same way, the *oh really* in (4), is responded to by an affirmative head nod from the prior speaker. In this respect the newsmark *oh really* can be seen as non-aligning with the prior speaker's ongoing activity and initiating a new action in the talk.

Interactional analysis of the 177 insertions in the collection resulted in the classification of 149 instances as 'alignment with activity in progress' and 28 instances as 'non-alignment'. While it is possible that the discrepancy in size of the two groups is a property of these particular meetings, an alternative explanation is that non-aligning actions which move away from the prior speaker's agenda are less likely to be accomplished in a short turn consisting of an insertion of just one intonational phrase. They may even include dispreferred actions that often require an elaboration or account of some kind within the same turn (cf. Schegloff, 2007, p. 63ff.).

3.2 Normalization

In order to address our hypothesis, we need reliable methods to measure the acoustic similarity of the insertion to the immediately preceding turn. Since F0 and intensity are likely to be prominent factors in the listener's perception of a prosodic match, we start by comparing the F0 contours and intensity of the adjacent turns. Our first step is to normalize the F0 range and the intensity range, in order to deal with cross-speaker discrepancies. 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, we also compare the F0 contours and weight the resulting similarity by intensity.

3.2.1 F0 normalization. As described in the introduction, 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 normalizations for each speaker are therefore needed. F0 normalization makes two comparisons possible: firstly of where each speaker locates their F0 contour within their overall range; and secondly of how far away from his or her mean F0 the speaker's contour falls or rises (i.e., the F0 span).

In order to normalize for F0, we computed the F0 of each of the three speakers over the entire length of every meeting. We used the YIN algorithm (de Cheveigne & Kawahara, 2002), which has the option to retain only those stretches of the F0 contour that coincide with high periodicity in the signal (by setting a threshold of aperiodicity, which here was set to 0.2).¹ The distribution of all F0 values obtained with this method shows multiple peaks: one peak corresponds to the values from the target speaker, and the others correspond to F0 values from the other speakers (see Figure 1). This phenomenon is due to crosstalk from the other two speakers that is picked up by the target speaker's microphone, which although low in sound level is able to influence the F0 distribution. 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 be much more intense than the crosstalk; since YIN identifies the most dominant pitch period in the acoustic signal, reliable F0 tracks can then be obtained. Summary statistics of the F0 values found in meetings B, C and D are shown in Table 1.

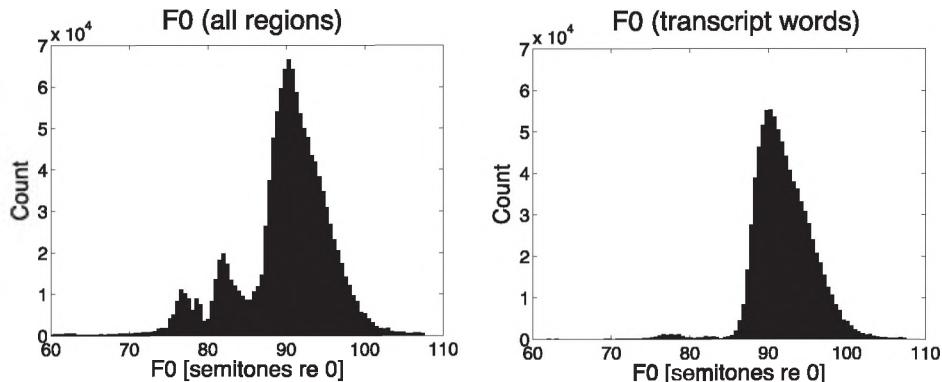


Figure 1. Distribution of F0 values for speaker B. Left panel: F0 distribution estimated from all regions of speaker B's speech, which is multimodal. The two peaks centered on 76 and 82 semitones are due to interference (crosstalk) from the two male speakers. Right panel: This problem was addressed by retaining F0 values only when the transcript indicated that speaker B was active. Note that the F0 distribution is skewed to the right.

Table 1. Summary of statistical measures of the distribution of F0 values in Hz and semitones (re 0 Hz).

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

The F0 distributions were skewed to higher F0 values, especially for the female speaker B (see Figure 1). The reasons for this effect are speculative; how speakers make use of their F0 span may depend on the environment in which the conversation takes place, the task in which the speaker is involved, and 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 might be different to the amount of excursion from the median of another speaker. In order to take this into consideration, we normalize the speaker's

distributions according to their standard deviation (cf. Heldner et al., 2010). Specifically, the F0 values were logarithmically scaled $\hat{f} = \log_2(F0)$, and then the normalized contour f 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, normalized for speaker characteristics, provide the basis for comparison across speakers with the aim of identifying prosodic matches.

3.2.2 Intensity normalization. Because the microphone channels might have been recorded with different levels, and because of uncertainty about the distance of the microphone from the mouth and individual differences in the intensity of speaking, normalization was also carried out for the intensity contours. First, an intensity contour is computed from the instantaneous power of the signal, smoothed according to the fundamental period (which is identified by the YIN algorithm, as described above). We then transform the intensity values into decibels and normalize them by subtracting the median value and dividing by the standard deviation.

3.3 Identification of prosodic matches

Our hypothesis is that in cases of matching of the two turns in the candidate matching pair, speaker B's F0 contour will match the domain of the prior speaker's contour. To claim that one contour matches some other contour, we need an objective measure of their similarity. We now describe such a metric, which takes into account both the movement of the F0 contour, and the range in which this movement occurs. Optionally, the measure of similarity may be weighted by the mean intensity of the two talkers; the motivation for this approach is explained below.

The metric for F0 similarity 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:

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

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 gives a similarity of zero. The parameter σ determines the width of the Gaussian function, and hence the tolerance of σ to F0 differences. Here we set $\sigma = 0.2$ by inspection. This value is not critical; qualitatively similar results were obtained across a range of σ values (see Section 4 below).

The equation given above describes how the similarity of two instantaneous F0 values can be quantified. However, our aim is to determine the similarity of two F0 contours, which will vary in length. Accordingly, we adopt the scheme shown schematically in Figure 2, in which sections of the F0 contour of speaker B's speech (f_B) are compared with sections of the F0 contour of speaker A's preceding speech (f_A). The length 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'.

Within the 3-second scope of the preceding speech, the F0 of the last intonational phrase (IP) of the prior speaker is compared with the F0 of the insertion IP. Because of differences in length of the prior speaker's IP and the insertion IP, sliding windows are used to perform the comparison. In practice, 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 length and the percentage of regions where one or other of the F0s is not available. To account for a

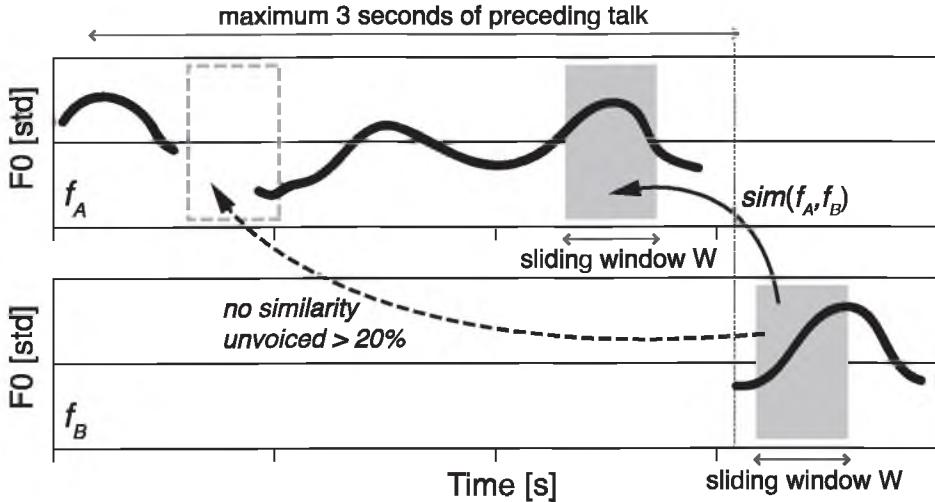


Figure 2. Scheme for measuring F0 similarity. Sections of the F0 contour for speaker A and speaker B, denoted f_A and f_B , are compared using a sliding window W to give a similarity value $sim(f_A, f_B)$. If more than 20% of the F0 values are missing within a particular window W (due to breaks in the F0 contour caused by unvoiced speech) then the similarity is not computed, as denoted by the dotted line and box.

reasonably long stretch of talk without introducing too many gaps in the F0 contours, we use a sliding window length 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, we compute the F0 similarity as

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

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 . Owing to the windowing over both F0 contours, sim_{AB} can be represented as a two-dimensional *similarity matrix* 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 gray scale.

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, because they are less likely to be corrupted by background noise or crosstalk from other speakers. The same can be argued from the 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.

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

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

where

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

Here, c is a constant added to ensure that the majority of $\alpha(t)$ values are positive (recall that the intensity values are in decibel units, so that they roughly correspond to perceived loudness and they have been normalized according to the median and the standard deviation of the intensity of each speaker). Occasionally values below c occur, and these are clipped to zero. The intensity-weighted F0 similarity metric (normalized by alpha) is then given by

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

The resulting similarity metric now tells us how similar the F0 contours are 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 makes a reduced contribution to the overall similarity.

We exemplify this from Extract (2) above. The talk of the first speaker (A) is followed by an “ah yeah” from the second speaker (B) and the action that the insertion does is to allow for continuation by the first speaker on his agenda. We hypothesize that the insertion “ah yeah” of speaker B is a prosodic match of the last part (the last IP) of speaker A’s utterance “I know where it is off”. We predict high similarity between the insertion and the end part of the preceding talk.

Figure 3 shows the F0 contours for the insertion “ah yeah” from speaker B and 3 seconds of the preceding talk from speaker A. For the same utterance pair, Figure 4 shows the individual intensity contours. The vertical dotted lines indicate the start and end of the IP of speaker A under investigation. Those parts of the utterances that best match each other are highlighted by thick black lines in both Figures 3 and 4.

Now we discuss the interaction between F0 and intensity in the current example. Speaker A’s F0 remains close to his median with little variation until towards the end of the utterance (“to Maplin”). The following “I kno*” is truncated, and is restarted with a higher F0 and also with a higher intensity. Over the “know” to “where” the F0 falls back to speaker A’s median. The insertion “ah yeah” starts one standard deviation above speaker B’s median, rises even higher to the syllable nucleus of “ah” and falls over “yeah” back even below her median. In intensity (Figure 4) the “ah yeah” of speaker B has two peaks, with the first peak on the first syllable higher than the peak on the second syllable. A similar two-peak structure can be seen in speaker A’s “I know” with the peak on “I” being higher than the peak on “know”.

Figure 5 shows a matrix of the computed similarity of the two F0 contours weighted by the mean intensity. The insertion of speaker B is represented along the ordinate, whereas the preceding talk of speaker A is shown on the abscissa. The more similar parts of the utterances are, the darker are the equivalent regions in the matrix. In the case of Extract (2) (C556), the highest similarity of the insertion after A’s stretch of talk is at 1.7 seconds in A’s talk. The similarity of the insertion grows from the start and is highest when it reaches 0.3 s. The maximum similarity at the cross section is 0.91.

Similarity scores were calculated for all pairs in our subset of the AMI corpus. As discussed above, some of these pairs were ‘action aligning’, while others were ‘action non-aligning’. For the first group, where we expected matches, we hypothesize that there will be higher prosodic similarity than for the second group, the expected non-matches.

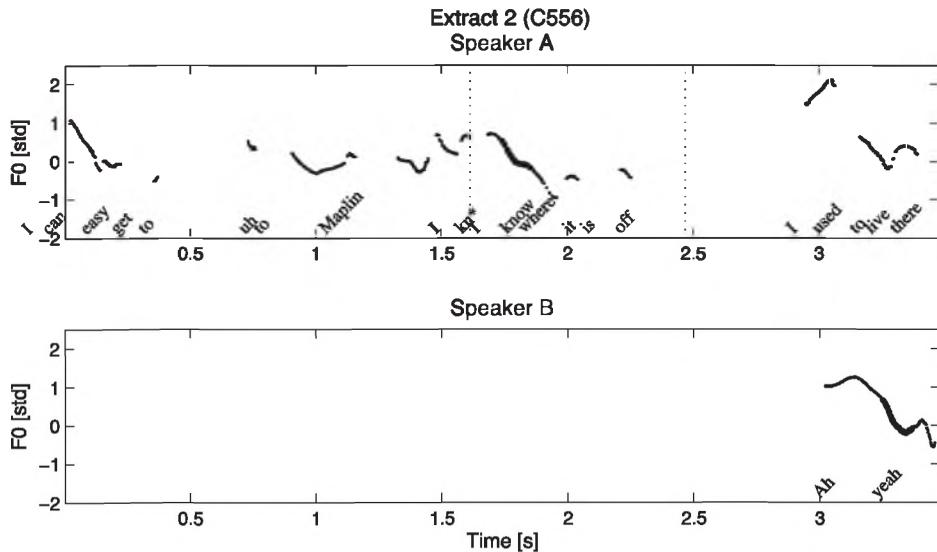


Figure 3. Speaker normalized F0 contours of the insertion “ah yeah” (bottom panel) and the preceding talk (top panel). The best matching parts (after intensity weighting) of the contours are highlighted by thick black lines. In this example (cf. Extract (2) (C556); lines 1 to 3) the final fall to the median of speaker B best matches the part of the F0 contour of the preceding talker which falls from above the middle of speaker A’s range (at 1.7 s).

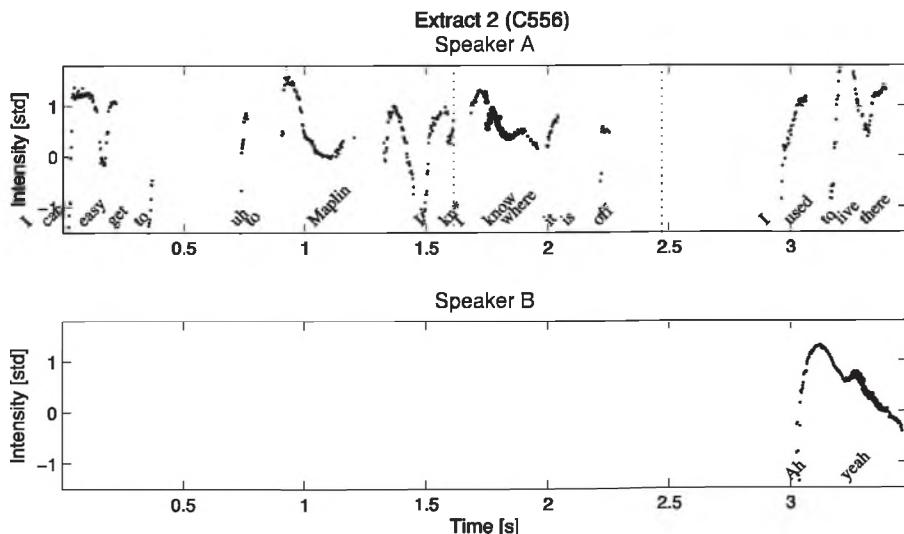


Figure 4. Speaker normalized intensity contours for both speakers. The contour of speaker B (bottom) shows two peaks on the two syllables of “ah yeah” (the first being higher than the second). A two-peak structure can also be identified in speaker A’s preceding talk “I know” (top), with the peak on “I” being higher than the peak on “know”.

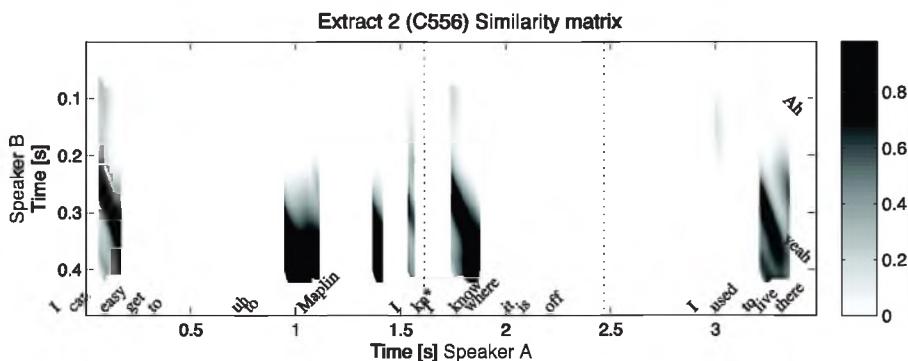


Figure 5. Similarity matrix of F0 contours modulated by intensity. The more similar the two contours are in the two parameters, the darker those parts in the matrix. Here, speaker B's utterance is most similar with speaker A's utterance at 1.7 s. The scale on the right indicates the strength of the similarity. The maximum between the two dotted lines is 0.91.

Table 2. Descriptive statistics (number, mean and standard deviation) of F0-only and intensity-weighted F0 similarity scores for alignments and non-alignments.

Interactional category	N	F0-only similarity		Intensity-weighted F0 similarity	
		Mean	Std. deviation	Mean	Std. deviation
Alignment	149	0.5893	0.3698	0.5798	0.3647
Non-alignment	28	0.3575	0.3689	0.3589	0.3690
Total	177	0.5526	0.3782	0.5449	0.3732

4 Results

Table 2 presents the similarity scores for the collection of insertions. Each similarity score represents the degree of similarity between the insertion, spoken by the female speaker B and the turn that immediately precedes it, spoken by one of the other speakers, A, C or D (see supplementary document for the result tables for all alignments and non-alignments: <http://las.sagepub.com/content/55/1/57/suppl/DC1>).

Of the 280 insertions in the collection, it was not possible to compute similarity measures for 103 because there were insufficient voiced sounds in the two stretches to be compared (see Section 3.3). The remaining 177 insertions have an overall mean F0 similarity score of 0.55 (F0-only similarity and intensity-weighted F0 similarity). For the insertions classified as aligning ($n = 149$), the mean F0 similarity without intensity weighting is 0.59 and with intensity weighting 0.58. For the insertions classified as non-aligning ($n = 28$), the mean F0 similarity is 0.36 both with and without intensity weighting.

The standard deviation is very high for both groups (between 0.37 and 0.38), which suggests a high degree of variation within both groups. The data are not normally distributed. One-tailed Mann-Whitney U tests (Mann & Whitney, 1947) were conducted to evaluate the hypothesis that the similarity score of ‘aligning’ insertions is higher, on average, than of ‘non-aligning’ insertions. The difference is significant for both the similarity scores based on F0 only ($z = -2.686, p = .003$) and the similarity scores based on intensity-weighted F0 ($z = -2.853, p = .002$), with a slight

Table 3. Results of one-tailed Mann-Whitney U tests for three values of the free parameter σ . For each value of σ , tests are reported for both versions of the similarity metric, F0 similarity weighted by intensity, and similarity based on F0 only.

F0 similarity	$\sigma = 0.1$		$\sigma = 0.2$		$\sigma = 0.3$	
	Intensity-weighted F0	F0 only	Intensity-weighted F0	F0 only	Intensity-weighted F0	F0 only
<i>z</i>	-2.923	-2.771	-2.853	-2.686	-2.719	-2.559
Exact sig. (1-tailed)	.002	.003	.002	.003	.003	.005

advantage for the latter. Aligning insertions have an average rank of 93.48 for F0-only similarity and 93.76 for intensity-weighted F0 similarity, while non-aligning insertions have an average rank of 65.18 and 63.70 respectively.

The results above were obtained for a specific value of the free parameter σ (0.2). Recall that this value determines the steepness of the function relating F0 difference to similarity score. Since the value of σ was set by inspection, it is important to determine the sensitivity of our analysis to this parameter. Accordingly, the analysis was repeated with σ values of 0.1 and 0.3. Similar results were obtained, as shown in Table 3, indicating that the similarity scores of ‘aligning’ insertions are significantly higher than those of ‘non-aligning’ insertions, for all values of σ tested.

5 Discussion

We assembled a collection of insertions, consisting of not more than one intonational phrase, produced by one speaker between turns produced by another speaker. From an interactional analysis, we classified them as either ‘aligning’ or ‘non-aligning’. On the basis of previous research, we developed the hypothesis that the ‘aligning’ insertions would be designed as prosodic matches to the immediately prior talk, while ‘non-aligning’ insertions would be designed as non-matches. In order to test this hypothesis, we developed an objective measure of prosodic similarity that could be applied to each ‘prior turn / insertion’ pair. This metric was primarily based on the similarity of the F0 contours of each pair, which were normalized for each speaker. As an additional condition, we also used a version of the metric in which instantaneous differences in F0 were weighted by intensity. The aim of the intensity weighting was to give a similarity measure closer to human perception, in which F0 differences that occur at low sound levels are less salient, and play a smaller role in determining the overall matching score.

The objective measure of prosodic similarity was applied to 177 insertion pairs, in order to address the following research question: *is pitch contour matching of an insertion and the immediately preceding turn used for alignment, and is non-matching used for non-alignment?* Applying the similarity index, we found a statistically significant difference between the two sets of ‘prior turn / insertion’ pairs, supporting the hypothesis that pitch matching is used for alignment, whereas non-matching is used for non-alignment. This provides the first objective acoustic demonstration, based on a substantial corpus of naturally occurring talk, that prosodic matching of pitch contours is both phonetically robust and interactionally relevant, as had been proposed by researchers in the phonetics of adult conversation such as Couper-Kuhlen (1996), Müller (1996) and Szczepek Reed (2006, 2009), as well as by Tarplee (1996) and Wells (2010) in the domain of child–carer interaction. It suggests that one source of phonetic orderliness in naturally occurring talk stems

from the requirement upon a next speaker to match the pitch contour of the prior speaker in order to demonstrate alignment with the talk in progress; or else to show non-alignment, for example in order to initiate a new action or direction, by demonstrably not matching the pitch contour of the prior speaker.

Although a significant difference was found, the results also showed a high degree of variability in similarity ratings across items within each class. There are a number of possible reasons why the degree of variability was so high. First, it is possible that some insertions were misclassified at the stage of interactional analysis; or indeed that the basis of the classification is mistaken in some way. For example, it may be that the dichotomous classification of insertions into ‘aligning’ and ‘non-aligning’ is too procrustean – that there is a wider range of possibilities that a second speaker can indicate, using additional prosodic devices to those of pitch matching vs. non-matching. This possibility is best investigated by further thorough interactional analysis of sequences of this type.

Second, it is likely that the acoustic measure of similarity can be refined and improved in order to reduce the number of ‘false positives’ (some pairs achieved high similarity scores although they are heard as non-matching) and ‘false negatives’ (where pairs which are heard as matching achieved a low similarity score). To our knowledge, this is the first attempt to measure the similarity of pitch contours in naturally occurring talk. It is a complex and challenging task. One area for improvement is in tuning the length of the comparison window to a domain that could represent the intonational phrase – without losing the comparability between examples. The latter could also be achieved by comparing how the prosodic features (extracted over time windows) develop over time in the candidate match pairs. We also note that there was little difference between the results obtained with the intensity-weighted F0 metric, and the metric based on F0 information only. It should not be inferred from this finding that intensity plays no role in prosodic matching. There are two possible reasons for this small difference: (1) the specific way in which intensity was used in our matching metric may not have been appropriate, (2) the acoustic signals from which the intensity contours were extracted are not optimal for this task: close-talking microphones are prone to record all sounds close to the mouth, including pops, smacks, breathing noises, etc. which could bias the intensity estimates.

Moreover, because the data are from spontaneous naturally occurring conversations they include extreme speaking styles such as fast speech and speech that is not intelligible to listeners and transcribers, where syllables and words are truncated, speech segments are modified and phenomena like devoicing and laryngealization are very frequent. Where voiced portions are still present, the number of fundamental periods is sometimes reduced to two or three (especially in the male voice), on which basis a fundamental period extraction or the analysis of its output can be problematic. This meant that a number of insertion pairs had to be omitted from the analysis. More widely however, it raises the question of what parameters the speaker might be matching. Szczepk Reed (2006) implies that almost any phonetic features may be implicated, including voice quality, for example. Thus our focus on F0 and intensity may be too narrow.

The instantaneous mean intensity across speakers was used here to weight the similarity of F0 contours, on the basis that F0 differences that occur at low sound levels are likely to be less perceptually salient (e.g., they are more likely to be masked by background noise in the room). However, it might be useful to compare the *overall intensity* between speakers’ turns as well, as continuers are often described as being quieter than the same lexical tokens used for other functions (Edlund, Heldner, Al Moubayed, Gravano, & Hirschberg, 2010).

6 Conclusions

Studies in the phonetics of conversation over the past thirty years have given rise to a wealth of insights and hypotheses as to how prosodic features are used for the purposes of interaction. With

developments in the recording and analysis of spontaneous talk it has become possible to test out such hypotheses by the analysis of large corpora. In the study reported here, this has been done using a publicly available corpus of spontaneous meetings talk which has the benefit of high quality recordings of the individual speakers on separate channels, as well as multiple video recordings.

The idea that we have explored is that pitch matching to the previous speaker provides the current speaker with a resource for demonstrating alignment with the prior speaker's action in progress; and that conversely non-matching provides a resource for non-alignment. This embodies the claim that choice of pitch contour is locally managed. The speaker does not have to refer to a lexicon or repertoire of intonation contours, to each of which a single meaning or set of meanings is associated. Instead, the current speaker designs the prosodic shape of the turn by reference to the prosodic shape of the preceding speaker's talk. If this finding is borne out by future research, it has quite radical implications for the modeling of prosodic features in applications such as automatic speech recognition and dialogue systems; and for understanding how children develop use of prosody (Wells, 2010), including atypical development such as the immediate and delayed echolalia found in cases of low-functioning autism (Local & Wootton, 1995). It may also have implications for how intonation is taught to second language learners.

Even though the results support the prosodic matching hypothesis, suggestions can be made for developing a more robust test of the hypothesis. These include possible improvements to the acoustic analysis techniques that could be borrowed from the speech technology community, such as more robust feature extraction and application of machine learning techniques such as hidden Markov models. At least as important, however, is the need for further detailed interactional research. Studies such as these require moving from qualitative CA analyses based on single cases to a quantitative classification, which runs the risk of oversimplifying the complexities and subtleties of conversation. These challenges point to the need for interdisciplinary collaboration in order to make further advances in the scientific analysis of spoken interaction.

Acknowledgements

This work was supported by the Marie Curie research training network "Sound to Sense" (grant number MRTN-CT-2006-035561). We are grateful to the members of the "Sound to Sense" network for their contribution to the development of this research; and to Elizabeth Couper-Kuhlen, Jens Edlund and Richard Ogden for their constructive comments on an earlier version of the article.

Note

1. The choice of pitch estimation is not crucial here, so long as it provides a means of specifying a threshold of voicing. We would expect similar results using other pitch determination algorithms, such as the one provided in Praat (Boersma & Weenink, 2011).

References

- 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.
- Boersma, P., & Weenink, D. (2011). Praat: Doing phonetics by computer [Computer program]. Retrieved from <http://www.praat.org/>
- Cooke, M. P. (1993). *Modelling auditory processing and organisation*. Cambridge, UK: Cambridge University Press.
- Couper-Kuhlen, E. (1996). The prosody of repetition: On quoting and mimicry. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: Interactional studies* (pp. 366–405). Cambridge, UK: Cambridge University Press.

- de Cheveigne, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111, 1917–1930.
- Drew, P. (1997). ‘Open’ class repair initiators in response to sequential sources of trouble in conversation. *Journal of Pragmatics*, 28(1), 69–101.
- Edlund, J., Heldner, M., Al Moubayed, S., Gravano, A., & Hirschberg, J. (2010). Very short utterances in conversation. In S. Schütz & G. Ambrazaitis (Eds.), *Proceedings from Fonetik 2010* (pp. 11–16). Lund, Sweden: Department of Linguistics and Phonetics, Centre for Languages and Literature, Lund University.
- French, P., & Local, J. (1983). Turn-competitive incomings. *Journal of Pragmatics*, 7(1), 17–38.
- 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 (Pragmatics & Beyond New Series*, 92). Amsterdam, The Netherlands: John Benjamins.
- 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.
- 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.
- Heldner, M., Edlund, J., & Hirschberg, J. (2010). Pitch similarity in the vicinity of backchannels. *Proceedings Interspeech 2010* (pp. 3054–3057). Makuhari: ISCA.
- Jefferson, G. (1984). Notes on a systematic deployment of the acknowledgement tokens ‘yeah’ and ‘mmhm’. *Papers in Linguistics*, 17(2), 197–216.
- Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed.), *Conversation Analysis: Studies from the first generation* (pp. 13–31). Amsterdam, The Netherlands: John Benjamins.
- Kaimaki, M. (2011). Transition relevance and the phonetic design of English call openings. *Journal of Pragmatics*, 43(8), 2130–2147.
- 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 (Eds.), *Where prosody meets pragmatics (Studies in Pragmatics 8)* (pp. 183–204). Bingley, UK: Emerald.
- 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: MIT Press.
- 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.
- 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.
- Moore, J., Kronenthal, M., & Ashby, S. (2005). *Guidelines for AMI speech transcriptions*. Switzerland: IDIAP; Edinburgh, UK: University of Edinburgh. <http://www.amiproject.org/>
- Müller, F. E. (1996). Affiliating and disaffiliating with continuers: Prosodic aspects of reciprocity. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: Interactional studies* (pp. 131–176). Cambridge, UK: Cambridge University Press.
- Pöppel, E. (2009). Pre-semantically defined temporal windows for cognitive processing. *Philosophical Transactions of the Royal Society B*, 364, 1887–1896.
- Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of ‘uh huh’ and other things that come between sentences. In D. Tannen (Ed.), *Georgetown University roundtable on languages and*

- linguistics, 1981. Analyzing discourse: Text and talk* (pp. 71–93). Washington, DC: Georgetown University Press.
- Schegloff, E. A. (1996). Turn organization: One intersection of grammar and interaction. In E. Ochs, E. A. Schegloff, & S. Thompson (Eds.), *Interaction and grammar* (pp. 52–133). Cambridge, UK: Cambridge University Press.
- Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in Conversation Analysis* (Vol. 1). Cambridge, UK: Cambridge University Press.
- Selting, M. (1996). Prosody as an activity-type distinctive signalling cue in conversation: The case of so-called ‘astonished questions’ in repair-initiation. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: Interactional studies* (pp. 231–270). Cambridge, UK: Cambridge University Press.
- Stensig, J., & Drew, P. (2008). Introduction: Questioning and affiliation/disaffiliation in interaction. *Discourse Studies*, 10(5), 5–15.
- Stensig, 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.
- Szczepk Reed, B. (2004). Turn-final intonation in English. In E. Couper-Kuhlen & C. E. Ford (Eds.), *Sound patterns in interaction* (pp. 97–118). Amsterdam, The Netherlands: John Benjamins.
- Szczepk Reed, B. (2006). *Prosodic orientation in English conversation*. Basingstoke, UK: Palgrave Macmillan.
- Szczepk Reed, B. (2009). Prosodic orientation: A practice for sequence organization in broadcast telephone openings. *Journal of Pragmatics*, 41(6), 1223–1247.
- Tarplee, C. (1996). Working on young children’s utterances: Prosodic aspects of repetition during picture labelling. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: Interactional studies* (pp. 406–435). Cambridge, UK: Cambridge University Press.
- Walker, G. (2004). *The phonetic design of turn endings, beginnings, and continuations in conversation* (PhD thesis). University of York, York, UK.
- Wells, B. (2010). Tonal repetition and tonal contrast in English carer-child interaction. In D. Barth-Weingarten, E. Reber, & M. Selting (Eds.), *Prosody in interaction* (pp. 243–262). Amsterdam, The Netherlands: John Benjamins.

Appendix A: Transcript symbols

hh	Audible outbreak; number of characters indicating the length in 0.1 s steps
.hh	Audible inbreath
:	Lengthening of preceding speech, for example a long hesitation “u::m”
(0.3)	Pause in seconds
(.)	Short pause (less than 0.1 s)
[Beginning of overlapping speech
*	Truncated speech at starts or ends of words, for example “ha*” or “*tion”
