Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 58 (2014) 11–34

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction

Céline De Looze [a,*], Stefan Scherer [b], Brian Vaughan [a], Nick Campbell [a]

[a] Speech Communication Lab, Trinity College Dublin, 7-9 South Leinster Street, Dublin 2, Dublin, Ireland
[b] Institute for Creative Technologies, University of Southern California, USA

Received 13 January 2012; received in revised form 27 July 2013; accepted 16 October 2013
Available online 30 October 2013

Abstract

Spoken dialogue systems are increasingly being used to facilitate and enhance human communication. While these interactive systems can process the linguistic aspects of human communication, they are not yet capable of processing the complex dynamics involved in social interaction, such as the adaptation on the part of interlocutors. Providing interactive systems with the capacity to process and exhibit this accommodation could however improve their efficiency and make machines more socially-competent interactants.

At present, no automatic system is available to process prosodic accommodation, nor do any clear measures exist that quantify its dynamic manifestation. While it can be observed to be a monotonically manifest property, it is our hypotheses that it evolves dynamically with functional social aspects.

In this paper, we propose an automatic system for its measurement and the capture of its dynamic manifestation. We investigate the evolution of prosodic accommodation in 41 Japanese dyadic telephone conversations and discuss its manifestation in relation to its functions in social interaction. Overall, our study shows that prosodic accommodation changes dynamically over the course of a conversation and across conversations, and that these dynamics inform about the naturalness of the conversation flow, the speakers' degree of involvement and their affinity in the conversation.

Keywords: Prosodic accommodation; Dynamics; Interactional conversation; Information exchange; Speakers' involvement and affinity

1. Introduction

Spoken dialogue systems make use of various language technologies and are increasingly being used to facilitate and enhance human communication, particularly through their use in Human Computer Interfaces (Oviatt, 1996; Coulston et al., 2002). These language technologies have use in a diverse range of fields including mobile communications (Ward and Nakagawa, 2002; Lu et al., 2011; Agarwal et al., 2011) internet search engines (Google, 2011; Apple Inc, 2011), games and assistive technologies developed for the elderly (Kleinberger et al., 2007) or communicatively impaired (Zhou et al., 2012). While these interactive systems can process the linguistic aspects of human communication, they are not yet capable of processing the important suprasegmental social information that is a pervasive part of human social interaction. Spoken interaction not only involves an exchange of propositional content but also the expression of affect, emotions, attitudes and intentions of the speakers. The ability for conversational partners to express, comprehend and react appropriately to these social signals is necessary for mutual understanding and successful communication (Boylan, 2004; Pickering and Garrod, 2006). Providing interactive systems with the capacity to process and exhibit these social signals will improve their efficiency and offer users more intuitive and appealing communicative interfaces.

Moreover, interactive technologies have been developed with the assumption that the user and the computer take

* Corresponding author.
E-mail address: deloozec@tcd.ie (C. De Looze).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.10.002

turns in a question-answer based interaction which fails to capture the complex dynamics involved in social interactions. Phenomena such as *accommodation*, *turn-takings*, *backchannels* and *overlaps* demonstrate coordination and adaptation on the part of interlocutors and suggest that conversation is not simply a start-stop interaction. Social interaction is a dynamic and joint activity where all participants are engaged and coordinate their behaviour in the co-construction of meaning (Mondada, 2001). It is an inherently complex activity as it requires a set of cognitive, linguistic and psychosocial skills, which allow individuals to understand each other and to establish a social relation. This *coordination* or *social resonance* (Tickle-Degnen and Rosenthal, 2007; Kopp, 2010) should ideally be implemented into automatic systems to make a robot or virtual agent a more socially competent interactant.

Developing automatic systems that are capable of recognising and understanding social cues and behaviours is, however, a difficult and ongoing process, as a complete understanding of social resonance has not yet been reached; in addition the development of these automatic systems requires the solving of unresolved issues related to social cue extraction, temporal and spatial alignment of extracted data as well as measurement and output representation and interpretation (Vinciarelli, 2009).

In particular, it has long been observed that conversational partners accommodate their pitch, intensity and timing behaviour to their interlocutors. However, at present, no automatic system is available to process interpersonal prosodic accommodation, nor do any clear measures exist that quantify its dynamic manifestation. Indeed, the majority of research has focused on its linear manifestation (over the course of an interaction). It is yet our hypothesis that it evolves dynamically with functional social aspects.

In this study, we test the assumption that prosodic accommodation is a dynamic phenomenon by investigating its evolution, at two levels of analysis – within and across conversations. We test the hypothesis that these dynamics are related to specific social functional aspects, looking at their correlation with crowd-sourced functional annotations regarding the perceived naturalness of the conversation flow, mutual understanding between speakers, speech interruptions, floor-holding patterns, speakers' degree of involvement and affinity.

We give hereinafter a brief review of the literature on prosodic accommodation and present the automatic system we developed for the measurement of its dynamic evolution.

2. Prosodic accommodation: forms, functions and dynamics

2.1. Forms and situational contexts

In many studies, it has been observed that, over the course of a conversation, speakers tend to accommodate their communicative behaviour to their interlocutor's and to the environment. A myriad of terms have been used to describe speakers' interpersonal adjustments, including convergence (Giles et al., 1991; Pardo, 2006), alignment (Pickering and Garrod, 2006), entrainment (Brennan, 1996), synchrony (Edlund et al., 2009), mimicry (Pentland, 2008) and chameleon effect (Chartrand and Bargh, 1999). The terms child-directed speech or motherese (Fernald et al., 1989) have also been employed to describe speakers' accommodation when talking to infants or children, foreign talk or foreignese (Ferguson, 1975; Zuengler, 1991; Smith, 2007) when interacting with non-native speakers and Lombard effect (Van Summers et al., 1988; Zeine and Brandt, 1988) when accommodating to a noisy environment. In this paper, the term *accommodation* is used in a generic way, covering all types of accommodation and defined as the way speakers adjust their speech to that of their interlocutor, *adapting* or *differentiating* it.

Speakers have been found to adapt their lexicon (Brennan, 1996; Pickering and Garrod, 2004; Nenkova et al., 2008), the grammatical and syntactic structure of their utterances (Levelt, 1982; Branigan et al., 2010; Cleland and Pickering, 2003; Haywood et al., 2005; Pickering and Ferreira, 2008), their pronunciation (Giles et al., 1991; Delvaux and Soquet, 2007; Babel and Bulatov, 2011; Aubanel and Nguyen, 2010; Bailly and Lelong, 2010; Pardo, 2006) and their prosodic characteristics to those of their partners (Natale, 1975; Gregory and Hoyt, 1982; Gregory et al., 1993; Stanford and Webster, 1996; Gregory and Dagan, 1997; Edlund et al., 2009; Levitan et al., 2011a; De Looze et al., 2011). They have been shown to display similar facial expressions (Bavelas et al., 1986; Hess and Blairy, 2001) and mimic gestures and body postures and movements (Condon and Sander, 1974; Meltzoff and Moore, 1977; Maurer and Tindall, 1983; Bernieri and Rosenthal, 1991; Chartrand and Bargh, 1999; Richardson et al., 2007; Shockley et al., 2007, 2009).

2.2. Prosodic realisation of accommodation

It has long been noted that conversational partners tend to exhibit similar pitch and intonation contours (Putman and Street, 1984; Giles et al., 1991; Zebrowitz et al., 1992; Gregory et al., 1993; Stanford and Webster, 1996; Gregory and Dagan, 1997; Collins, 1998; Shepard et al., 2001; De Looze et al., 2011), voice intensity level (Black, 1949; Meltzer and Morris, 1971; Natale, 1975; Gregory and Hoyt, 1982; Coulston et al., 2002; De Looze et al., 2011), speech rate and speech timing (Matarazzo and Wiens, 1967; Webb, 1972; Welkowitz and Kuc, 1973; Street et al., 1983; Woodall and Burgoon, 1983; Giles et al., 1991; Jaffe, 2001; Kousidis et al., 2008; McGarva and Warner, 2003; Edlund et al., 2009; De Looze et al., 2011).

For instance, Collins (1998) and Stanford and Webster (1996) observed global pitch level adaptation (in terms of meanf_0), using unconstrained conversations and interviews of English. Similarly, we recently found evidence of pitch level and pitch range adaptation (in terms of medianf_0 and sdf_0), using English unconstrained and task-based dia-

logues (Vaughan, 2011; De Looze et al., 2011; De Looze and Rauzy, 2011). Heldner et al. (2010), Levitan et al. (2011a) and Levitan and Hirschberg (2011) have also observed local aspects of pitch adaptation, investigating pitch entrainment in backchannels and in speech preceding backchannels. Using spontaneous dyadic conversations, where native speakers of American English played a series of computer games, they have shown that speakers adapt the pitch of their backchannels to the pitch of their interlocutors' preceding utterance as well as come to use similar backchannel-preceding cues over the course of the conversation.

Natale (1975) has also observed mean vocal intensity adaptation in English non-directive interviews. Recent studies by Kousidis et al. (2009), Vaughan (2011), De Looze et al. (2011), De Looze and Rauzy (2011), Heldner et al. (2010) and Levitan et al. (2011a) further confirmed these findings, where adaptation in mean-Intensity and sd-Intensity has been observed both at global and local levels of task-based and unconstrained dialogues of English and Swedish.

Regarding speech rate and speech timing, Matarazzo and Wiens (1967) found for instance evidence of pause duration adaptation between interviewer and interviewee: the pause duration of the interviewee was directly influenced by those of the interviewer (Matarazzo and Wiens, 1967). This was recently observed in spontaneous conversations of English and Swedish (De Looze et al., 2011; De Looze and Rauzy, 2011; Edlund et al., 2009). McGarva and Warner (2003) also examined vocal activity rhythm accommodation in conversational dyads and found such adaptation in some but not all of the interactions; over a third of the conversations had moderate to significant levels of vocal activity adaptation. Mc Garva hypothesizes that this may be due to an underlying chronobiology and the need for vocal activity adaptation, in some cases, to build up over time rather than happen spontaneously.

2.3. Functional role in social interaction

Adaptation is a particularly important aspect of social interaction as it facilitates, through the alignment of cognitive representations, comprehension and understanding between interlocutors. It correlates with the communicative success of the interaction, by decreasing misunderstandings and attaining goals faster (Boylan, 2004; Pickering and Garrod, 2004; Parrill and Kimbara, 2006; Pickering and Garrod, 2006). In particular, in human–robot interaction (HRI), human users' adaptation improves the content-information exchange and hence, the conversation flow (Breazeal, 2002; Branigan et al., 2010). In addition, adaptation participates in increasing the social success of the interaction in terms of rapport (i.e. harmonious relation and mutual attention) and affiliation (Tickle-Degnen and Rosenthal, 2007; Lakin and Chartrand, 2003; Pickering and Garrod, 2006; Shepard et al., 2001; Miles et al., 2009; Kopp, 2010).

In particular, research has suggested that prosodic adaptation is a subconscious method of achieving social approval and acceptance and is utilised to identify with a particular social group (Matarazzo and Wiens, 1967; Giles et al., 1991; Chartrand and Bargh, 1999). According to Chartrand and Bargh,"*the chameleon effect operates in a passive, non-goal dependent manner to create greater liking and ease of interaction*" (Chartrand and Bargh, 1999, p. 901). Natale (1975) have for instance investigated speakers prosodic accommodation according to their degree of social desirability using Crown and Marlowe's social desirability test (Crowne and Marlowe, 1960). They have found that individuals who scored higher in terms of social desirability were more likely to adapt their voice intensity level and timing patterns to those of their partners in contrast to those with a low social desirability score.

Prosodic accommodation can be influenced by the participants' social background. Giles et al. (1991) explain for instance that the participant of lower social status would tend to adapt more to the speech of the participant considered to be of a higher social status. According to Communication Accommodation Theory (CAT; Giles et al., 1991), individuals accommodate to their partners on an adaptation-maintenance-differentiation continuum, where at the extreme other end, they differentiate their behaviour to that of their interlocutor.

In a previous experiment (De Looze et al., 2011), we have also observed that the amount of prosodic accommodation (in terms of pitch range and voice intensity level) displayed by interactional partners is correlated to their involvement – or active engagement, interest in taking part in the conversation; the more involved they are, the more they adapt their prosodic variations.

In addition, individuals who adapt to their partners have been evaluated more positively than those who do not (Giles et al., 1991), in particular in terms of power, attractiveness and intelligence (Gallois and Callan, 1988, 1991). Investigating accommodation in relation to positive and negative attitudes in married couples' problem-solving interactions, Lee et al. (2010) have found, that high levels of adaptation in pitch are correlated with a positive attitude while high levels of differentiation with negative attitude.

In the context of human–computer interaction (HCI), accommodation has also been found to be an important aspect of the interaction. Users' amicability for a machine increases when it adapts to their prosody (Suzuki and Katagiri, 2007). Ward and Nakagawa (2002) found for instance that a telephony system that adapts its speech rate with the users' is rated more favourably than those that do not.

These findings regarding prosodic accommodation in HC and HRI (Oviatt, 1996; Bell et al., 2003; Coulston et al., 2002; Breazeal, 2002) suggest that prosodic accommodation is such an important, constituent part of vocal social interaction, as well as being a largely unconscious process, that it manifests automatically, regardless of the

interlocutor (a human vs. a machine) or the conversation aim (task-based vs. chat). According to Lakin and Chartrand (2003), accommodation would have become automatic over the course of human evolution, playing an important role as a necessary pre-requisite for communicating and for maintaining harmonious relationships within a group. It would have evolved to act as a social glue, creating, facilitating and enhancing social links between individuals.

2.4. Measurement and quantification

While prosodic accommodation is a ubiquitous component of social interaction, its modelling is a difficult task. At present, no automatic system is available to process conversational partners' prosodic accommodation, nor do any clear measures exist that quantify this phenomenon adequately. In particular, the metrics developed so far failed to capture its dynamic manifestation.

2.4.1. Capturing accommodation dynamics

Numerous studies have examined speech accommodation with the assumption that it is largely a linear phenomenon, with accommodation usually increasing over the course of a conversation. Burgoon et al. (1995) define adaptation as "*the situation where the observed behaviours of two interactants, although dissimilar at the start of the interaction, are moving towards behavioral matching*".

In these studies, increased similarity in prosody over time has been assessed by comparing conversational partners' degree of prosodic adaptation for the first and second halves of the conversation or for its first, second and third parts (Jaffe and Feldstein, 1970; Suzuki and Katagiri, 2007) as well as across several conversations (Natale, 1975). If prosodic adaptation was shown to increase in the second and third parts of the conversation, or in the second and third conversations, this was taken as evidence that speakers' prosody has become more similar over the course of the conversation or across conversations. Natale (1975) examined, for instance, the amount of vocal intensity adaptation in three 10-min extracts of interview interactions that were separated by one-week intervals. Measuring intensity convergence as the absolute difference in mean and standard deviation intensity between speakers, he found that the level of adaptation was greater in the second and third interactions than in the first.

However, what makes a conversation an interactive dialogue, are the dynamic changes involved in spoken interaction. Interlocutors do not remain involved to the same degree over the whole course of a conversation; as they may change from being inactive to talking, going through phases such as listening, thinking, arguing a point or giving feedback. It can thus be assumed that accommodation undergoes similar dynamic changes in real-life spoken interaction. Earlier and recent work in this area has confirmed this assumption (Stanford and Webster, 1996; Gregory and Dagan, 1997; Collins, 1998; Levitan and Hirschberg, 2011; Kousidis et al., 2008; Edlund et al., 2009; Vaughan, 2011; De Looze and Rauzy, 2011; De Looze et al., 2011). In previous experiments (Vaughan, 2011; De Looze et al., 2011; De Looze and Rauzy, 2011), we found for instance that the amount of prosodic accommodation (in terms of meanf_0, sd-f_0, mean-Intensity, sd-Intensity and pause duration synchrony and asynchrony), changes several times over the course of unconstrained and task-based interactions of English.

In Levitan and Hirschberg (2011) prosodic accommodation was reported to be both a linear and a dynamic phenomenon, where it was measured over the course of a whole conversation (described in their paper as the conversation level) and at the turn-level. For the conversation level analysis, each conversation was split into two parts, inferring a linear manifestation of adaptation when the differences between the prosodic values of the two speakers was less in the second half. The turn-level analysis indicated that speakers match their interlocutors at these turn exchanges, thus attesting to the dynamic manifestation of adaptation.

Due to the fact that prosodic accommodation has been shown to increase continuously as well as to vary over the course of the conversation, we assume that speech accommodation can actually manifest as both a linear and dynamic phenomenon. Some cues may exhibit a largely linear rate of accommodation, increasing or decreasing in a linear fashion over the course of an interaction or across several consecutive interactions. Likewise, accommodation between some parameters can fluctuate over the course of a conversation or across several conversations.

We assume that the linear and dynamic manifestation may depend on a number of factors. In a form-function mapping, we hypothesize that when the functional aspect of accommodation is static over the course of a conversation, its forms are manifested in a linear trend. One may for instance observe a linear manifestation of accommodation in situations where interlocutors adapt their partners' speech style or regional accent. Accommodation may increase over time as speakers get used to and assimilate their interlocutor's accent. On the other hand, if it is linked to speakers' social intentions, we would expect a dynamic trend where accommodation fluctuates with the conversational partners' change of states. One may also observe a dynamic manifestation of accommodation as it fluctuates with the speakers' degree of involvement in a conversation (De Looze et al., 2011; De Looze and Rauzy, 2011).

While the metrics developed so far are capable of measuring the linear increase of accommodation, they however fail to capture or consider its temporally dynamic nature. Further work is thus needed.

2.4.2. Prosodic cues extraction and time-alignment

Measuring prosodic accommodation is not an easy task. The difficulty one encounters when measuring prosodic accommodation is that it is not time-aligned. Speakers do not accommodate to each other immediately due to the

inherent temporally reactive nature of conversational speech. It is therefore important to find a way to meaningfully compare prosodic features between multiple interlocutors and to ensure they are time-aligned to enable detection and measurement of accommodation.

Current approaches comprise two types of methods: utterance or turn-level-based (e.g. Levitan and Hirschberg, 2011) and time aligned moving average (TAMA) methods (Kousidis et al., 2008). Fig. 1 shows the differences between the two approaches. The red rectangle (analysis window) refers to the analyzed audio snippets. The third approach, i.e. a HYBRID utterance sensitive approach, is a trade off between the two extremes and shall be discussed in Section 3.2.

2.4.2.1. Utterance/turn-level-based methods. The utterance or turn-level based approaches analyse prosody within two consecutive utterances or turn-levels, each spoken by different speakers. The advantage of such a fine-grained method is that it takes into account speakers' vocal activity rhythm. In Heldner et al. (2010) and Levitan et al. (2011a), this has proved to be efficient to capture prosodic accommodation in backchannels in natural conversations. Investigating accommodation at the utterance or turn-level (e.g. as in Levitan and Hirschberg, 2011) however includes the assumption that it is a local phenomenon only and that the effect of a partner's speech characteristics on his/her interlocutor's is found immediately after each utterance. In this view, conversations are reduced to a ping-pong interaction, where speaker B's utterance is only linked to speaker A's preceding utterance. This over-simplifies what really takes place in real-life conversations. Real-life interactions are not a composition of one-question-one-answer dialogues. The effect of a speakers speech characteristics on his/her partners may be found after some temporal delay, which may exceed the utterance or the turn domain. A wider temporal span seems therefore also worth investigating for the study of prosodic accommodation. This could be done by comparing several utterances as was done in Nishimura and Kitaoka (2008).

2.4.2.2. The TAMA method. In contrast to utterance/turn-level-based approaches, the TAMA (time-aligned moving average) method proposed by Kousidis et al. (2008) analyses the audio in a fixed window, averaging values out over the duration of the window. For window-based approaches, it is crucial to find the window or temporal span that is most suitable to extract prosodic features and allows for a comparison between the partners. In particular, this method is based on the extraction of average prosodic values for each speaker from a series of overlapping fixed length windows (frames). Speech intervals are thus cut at the windows' boundary regardless of the size of the utterance. Fig. 1 (middle chart) shows the moving window along speakers' interaction (represented by a conversation chart). Large and overlapped frames give a smoothed contour for the prosodic parameter being analyzed, while short frames detect more abrupt modifications

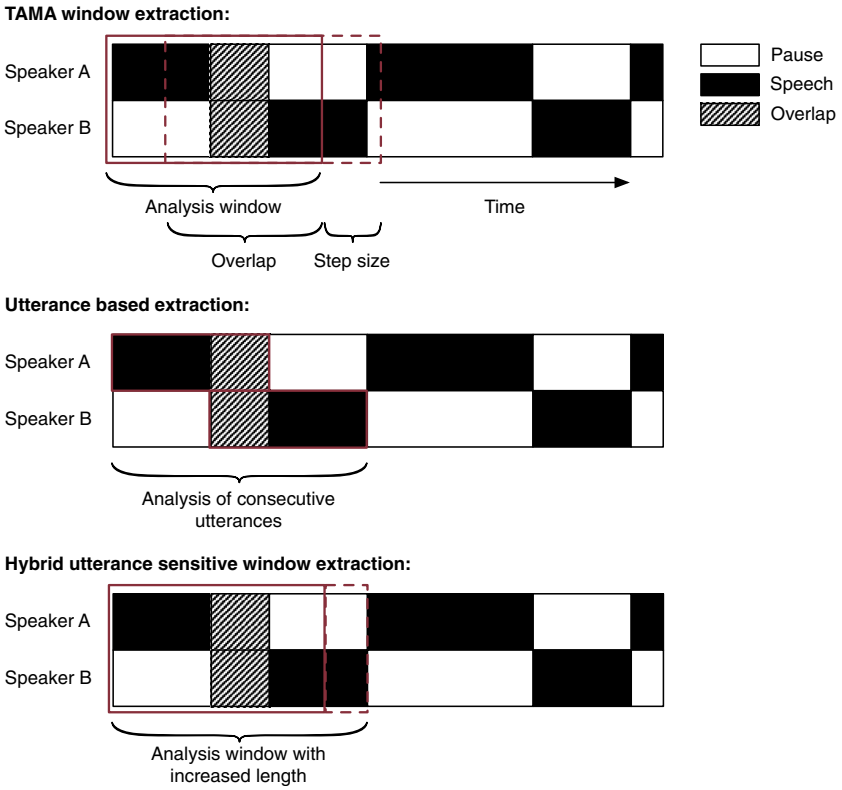

Fig. 1. Illustrating comparison of the three analysis methods, i.e. time aligned moving average (TAMA) based, utterance based, and utterance sensitive window based.

(Kousidis et al., 2008). The major drawback however is that utterances are randomly cut, i.e. even if a speaker has not yet finished talking. This method can also be problematic if, within the analysed frame, only one speaker is talking. This means that no meaningful values can be calculated for the other speaker and thus accommodation measurements may be skewed. One method of resolving this is to use interpolation to generate values from the previous and succeeding frames. However, one is limited to using small window sizes with large overlaps as faulty values may result if a large window size with only a small amount of overlap is used in combination with interpolation. These problems can be addressed through the use of a hybrid approach that is sensitive to utterance boundaries (discussed further in Section 3.2).

Developing a system for the automatic measurement of prosodic accommodation not only requires the ability to accurately capture and measure its dynamic manifestation, but also requires the defining of an appropriate temporal span for prosodic cue extraction and prosodic measurement. In the following section, we present the prosodic accommodation dynamics (PAD) tool we developed, that takes the above considerations into account.

3. The PAD (prosodic accommodation dynamics) tool

3.1. States of accommodation

In order to capture the different manifestations of prosodic accommodation, we have proposed, in De Looze and Rauzy (2011), a set of states of accommodation (cf. Fig. 2) based on the three categories – *adaptation*, *differentiation* and *maintenance* – determined by Giles et al. (1991) in their Communication Accommodation Theory. Adaptation refers to the tendency of conversational partners to accommodate their communicative behavior to each other throughout spoken interaction so as to become more similar. Differentiation, on the contrary, is their tendency to exaggerate their differences. Maintenance is the situation when neither conversational partner is affected by the other's communicative behavior.

In our definition, adaptation and differentiation are each described according to two underlying phenomena. Adaptation in terms of *synchrony* and *convergence*; differentiation in terms of *asynchrony* (or *symmetrical synchrony*) and *divergence*. Following Edlund et al. (2009), synchrony refers to the situation where two speakers exhibit temporally or simultaneously similar prosodic behaviours (e.g. due to linguistic, paralinguistic factors). This means for instance that when a speaker raises his voice intensity, his interlocutor does it too. Convergence, on the other hand, is realized when conversational partners' behaviours begin to accommodate toward a common point or prosodic matching (up to a certain point of equilibrium defined by physiological, cognitive, functional or social constraints). This means that in the situation where speaker A's speech rate is fast and speaker B's speech rate is slow, speaker A slows down his speech, speaker B accelerates it, so that they use a common speech rate. Asynchrony is the tendency for speakers to differentiate their prosodic variations from the other's, with variations resulting in mirror or symmetric patterns. Divergence refers to the tendency to move apart in different directions.

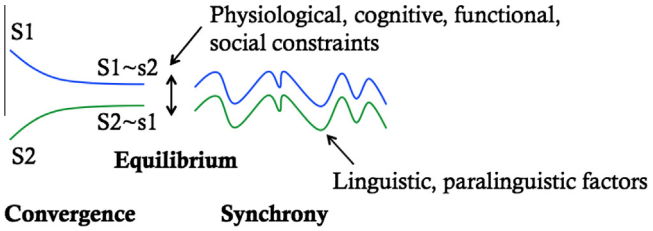

Fig. 2. Schematic representation of convergence and synchrony.

It is assumed that these underlying phenomena can be exhibited individually or in combination, resulting in 7 possible different states:

1. Three states of adaptation:
   (a) synchrony (synchrony/maintenance),
   (b) convergence (maintenance/convergence),
   (c) both synchrony and convergence (synchrony/convergence).

2. Three states of differentiation:
   (a) asynchrony or symmetrical synchrony (asynchrony/maintenance),
   (b) divergence (maintenance/divergence),
   (c) both asynchrony and divergence (asynchrony/divergence).

3. State of maintenance:
   (a) no adaptation and no differentiation (maintenance).

In this paper we focus on the measurement of the *dynamics of prosodic synchrony*.

3.2. Prosodic cues extraction

In its earliest version (De Looze and Rauzy, 2011; Vaughan, 2011), our tool made use of the TAMA method as proposed by Kousidis et al. (2008) to extract prosodic parameters. In the current version of the system, we use a HYBRID method based on utterance-based and TAMA methods (Fig. 1). Instead of randomly cutting the speech of the speakers, the moving windows are extended to the start and end of the utterances at the left and right boundaries of the window. In particular, this means that average values of prosodic cues are automatically extracted from a series of overlapping windows (frames) of a default-fixed length which are extended to the utterance temporal span at the window boundaries.

Such a method therefore allows both the consideration of speakers' vocal activity rhythm and speaker-time-aligned prosodic cue extraction. The argument for an utterance-sensitive system is that the functional aspects of prosodic accommodation may not change within an utterance but rather between utterances. The prosodic features, extracted from the entire utterance, are therefore representative of the utterance prosody and its functions in the interaction. With such a system, different window sizes and time steps can also be tested. This means that prosodic accommodation can be investigated at different temporal spans, which can offer insights of the hierarchy and embedment of prosodic accommodation at different levels of the interaction.

3.3. Prosodic measurements

The tool extracts a set of different acoustic parameters using the phonetic software Praat (Boersma and Weenink, 2006) and the Matlab signal processing software. These parameters account for pitch range, articulation rate and voice intensity.

- Pitch range: fundamental frequency median f_0 and standard deviation about the mean (sdf_0). The medianf_0 and sdf_0 are given on a logarithmic scale, the octave scale (i.e. $log2(Hertz)$) in order to facilitate gender comparisons. In order to avoid possible pitch tracking errors, pitch floor and pitch ceiling (when creating a Pitch Object) were set to the values $p_{15} \cdot 0.83$ and $p_{65} \cdot 1.92$ respectively, where p_{15} and p_{65} denote the 15th and 65th percentile respectively (De Looze, 2010).
- Voice intensity: standard deviation of intensity (sdInt) and its median (medianInt).
- Articulation rate: number of syllable nuclei per second (syllsec).

Speech/silent intervals and syllable nuclei are automatically annotated. Speech/silent intervals are detected using a method based on long-term modulation spectrum energy features (Maganti et al., 2007). Detection of syllable nuclei is performed using the method introduced in De Jong and Wempe (2009), which is based on intensity peak detection of voiced segments of speech.

Fig. 3 shows an example of the various feature values obtained for each moving window for a conversation between two speakers. This enables us to visually investigate the evolution of prosodic accommodation.

3.4. Prosodic synchrony measurement

To measure the synchrony in the development of the extracted parameters for each interlocutor, we utilized the standard Pearson correlation coefficient $\rho_{xy} \in [-1, 1]$, that measures linear dependencies between two sets of observations x and y (belonging to the two separate interlocutors):

$$\rho_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)\sum_{i=1}^{N}(y_i - \mu_y)}{(N-1)s_x s_y}, \quad (1)$$

where $|x| = |y| = N$, μ_x the mean value of x (respectively μ_y), s_x the standard deviation of x (respectively s_y), and $x_i \in x\ \forall i = 1, \ldots, N$ (respectively y_i). For large $\rho_{xy} \gg 0$ we have strong linear dependencies, which indicates a synchronous behavior of the prosodic parameters over the analyzed fragment. Small values $\rho_{xy} \ll 0$ indicate strong asynchronous developments of the observed parameters. For values close to zero no linear correlation is observed, i.e. a state of maintenance is present.

As Eq. (1) requires the same number of observations in both observation sets x and y, we utilize a temporal windowing as explained in the following. In this work, $N = 10$ and therefore the threshold for significant positive or negative correlations for $\rho_{xy} \approx \pm.5$ at a significance level of $p < .05$.

3.5. Temporal span

In order to investigate the dynamics of prosodic synchrony within and across conversations, the Pearson correlation synchrony analyses are executed on multiple levels of granularity, i.e. sub-conversation, conversation, and supra-conversation levels.

For the within conversation (sub-conversation) analyses of synchrony, we group 10 windows of the HYBRID feature extraction with a step size of 5. This means that prosodic accommodation strength is calculated for a period of 100 s for every 50 s (cf. Fig. 3 for an example result).

For the analyses of accommodation across conversations, we calculate the ratios (or percentages) of states of synchrony/asynchrony. The time in the various states is normalized to the total length of a conversation to give the ratio; ratios add up to 1.

Prosodic synchrony is also computed for the first and second halves of the conversations as well as for the first, second and third parts of the conversations in order to investigate whether prosodic accommodation continuously increases over the course of the conversation, as was reported in the literature.

3.6. Real vs. pseudo-interactions

In order to investigate whether the moments of synchrony are meaningful and are not capturing accidental or coincidental phenomena, we use a method similar to Ramseyer and Tschacher (2010) and Ward and LItman (2007); our model creates a number of artificial conversations and from them computes pseudo-synchrony coefficients. These coefficients are then compared, using both a z-score transformation procedure and a Mann–Whitney U test, with those obtained from real conversations. If synchrony is significantly higher for real than for pseudo-interactions, this is taken as evidence that the prosodic

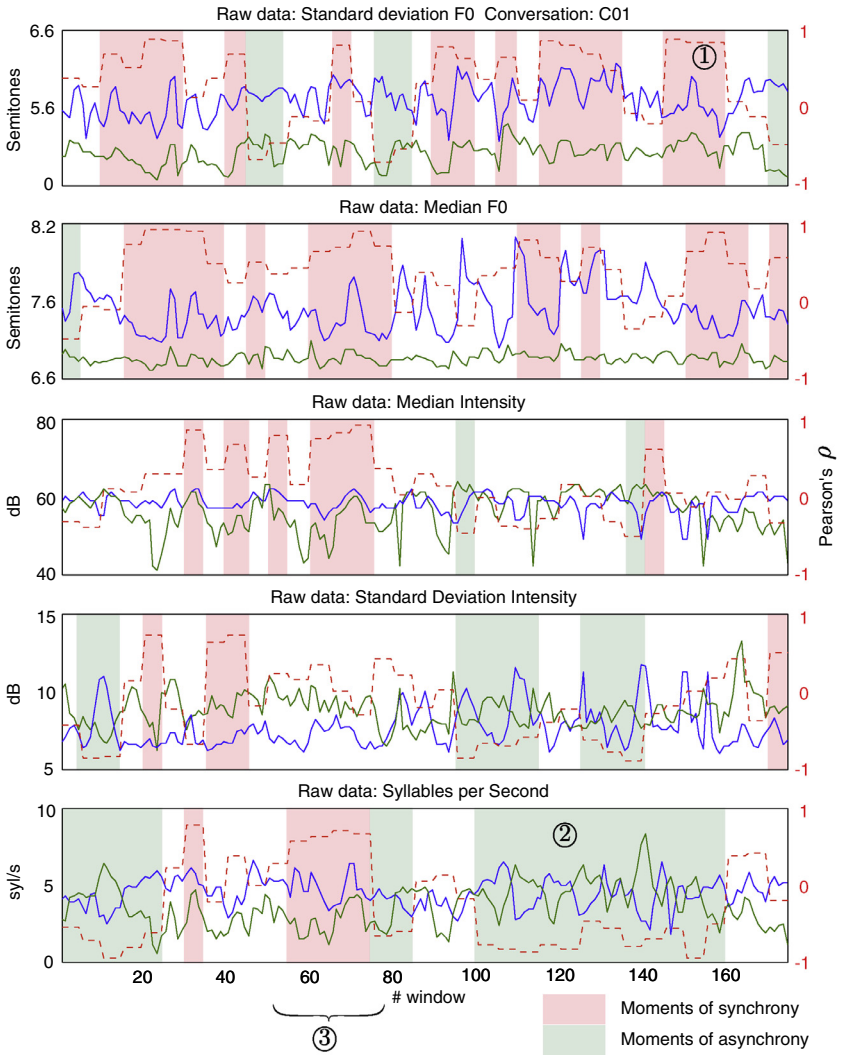

Fig. 3. Raw values of Pair 1's first conversation of each extraction window for each speaker (Speaker 1 in blue; Speaker 2 in green) for the respective feature (left ordinate axis) along with Pearson's ρ values (dashed red line; right ordinate axis). Phases of synchrony are highlighted in red phases of asynchrony in green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

synchrony captured in our data is not random or accidental: essentially background prosodic accommodation 'noise'.

Since there were 41 real conversations, there were 82 conversational 'sides' from which we created the pseudo/fake conversations. Having a total of 82 conversational side gave us a total of 3321 possible combinations without repetitions N, with

$$N = \frac{n!}{(n-r)!r!} \tag{2}$$

where n is the number of variables (in this case conversational sides) and r is the total number of variables in each possible combination. Fake conversations were then analysed and correlation measures computed as per the analysis of the real conversations. After discarding some fake conversations that were unusable due to computational errors, we were left with a total of 2966 fake conversations. This method meant that some of the fake conversations were made up of the same speaker from different conversations but since these conversations never took place, we made the assumption that any moments of accommodation would be accidental and not related to any functional aspect: similar to any non-functional moments of accommodation between two people with a similar speaking style and similar prosodic characteristics.

4. Dataset

In this study we used a subset of a corpus of recorded telephone conversations (Campbell, 2004) of natural daily speech designed to better understand changes in speech styles and attitudinal aspects of speech in real-life conversa-

tions. Real-life conversations provide richer social dynamics than task-based dialogues. In total the corpus comprises more than 50 h of unconstrained spontaneous Japanese speech.

From over one hundred 30-min telephone conversations (approximate duration), we chose 40 conversations from 6 speakers (three men and three women), which were recorded over a period of several months. The speakers, all native Japanese, met once a week to talk with specific partners in a separate part of the same building over an office telephone line. They did not see their partners or socialize with them outside of the recording sessions. They were initially strangers to each other, but became better acquainted over the period of the recordings (i.e. 10 conversations for each pair). There were no constraints on the content of the conversations other than that they should last for thirty-minutes. The pairs we chose include all three gender combinations (female–male, female–female, and male–male). Further, two speakers were included in two separate pairs. A complete list of conversations is presented in Table 1. While talking, speakers wore a head mounted Sennheiser HMD-410 close-talking dynamic microphone and recorded their speech directly to digital audio tape at a sampling rate of 48 kHz. The mono-recordings were subsequently down-sampled to 16 kHz and synchronized to facilitate a time-aligned analysis.

5. Experiment 1: evolution of prosodic synchrony

5.1. Objectives and hypotheses

In this first experiment, we investigated the evolution of prosodic synchrony at two levels of analysis – within and across conversations. In particular, we investigated whether prosodic synchrony continuously increases or dynamically varies over the course of a conversation and across several consecutive interactions. According to what has been reported in the literature, it could be expected that speakers' prosodic characteristics become more and more similar over the course the conversation, and across several conversations, in particular herein, where speakers' acquaintance increases over the course of the recordings. We assume however that, because of the dynamic nature of social interaction, prosodic accommodation changes variably throughout the course of a conversation and across conversations, resulting in several phases of synchrony, asynchrony and maintenance.

Table 1
List of the four pairs' conversations analysed in this study.

Pair ID	ID Speaker 1	Gender Sp. 1	ID Speaker 2	Gender Sp. 2
Pair 1	FB	Female	FC	Female
Pair 2	FA	Female	MA	Male
Pair 3	MA	Male	MB	Male
Pair 4	MB	Male	MC	Male

5.2. Method

The PAD tool, described in Section 3, was used in this experiment. Prosodic cues were extracted using the HYBRID method. The default window size was set to 20 s and the time step to 10 s, then extended to the start and end of the utterances at the left and right boundaries of the window; prosodic cues were therefore extracted for each speaker about every 10 s.

As previously mentioned, utterances were automatically annotated using a method based on long-term modulation spectrum energy features, for which speech intervals and silent intervals were detected. On average, for all conversations, mean utterance duration is of 9.3 s, with a standard deviation of 2.92 and mean silence duration is of 11.29 s, with a standard deviation of 2.9. With this utterance extension, window lengths ranged from 20 s to 26 s; mean window length of 20.60 s with a standard deviation of 0.68. In each HYBRID window, speech and silent intervals were distributed as follows: in average, speech intervals represented 46.52% of the window length, with a standard deviation of 14.63, while silent intervals represented 56.42% of the window length, with a standard deviation of 14.53.

Prosodic values were calculated for speech intervals only, silent intervals being excluded from the calculation. They were calculated proportional to the utterances (or speech intervals) length within the window, i.e. they correspond to weighted means.

Synchrony dynamics within conversations were investigated using three different temporal spans: synchrony coefficients were computed for the halves and thirds of the conversations as well as for the moving windows of 110 s-length (with an overlap of 50 s); the evolution of synchrony across conversations, by calculating the amount of synchrony, asynchrony and maintenance for each conversation.

5.3. Results

5.3.1. Synchrony dynamics within conversations

5.3.1.1. Conversation halves. Paired t-tests were first carried out to test whether synchrony coefficients obtained for all prosodic parameters in the first part of the conversation are significantly different from those of the second part. Significance level was set at $p < .05$.

Results reveal no significant difference of synchrony degree between the first half and the second half of the conversations for all prosodic parameters, except for $\mathrm{sd}f_0$ where the synchrony coefficient is found to be lower in the second part of the conversations ($p = 0.00535$).

5.3.1.2. Conversation thirds. T-tests were also carried out to investigate whether synchrony coefficients obtained for all prosodic parameters are significantly different in the first, second and third parts of the conversation. Significance level was set at $p < .01$.

As for conversation halves, no significant difference in the synchrony degrees was found between the beginning, mid and end parts of the conversation for all prosodic parameters.

Summary. These results show that prosodic synchrony does not continuously increase over the course of the conversation. They suggest it rather varies dynamically. The analysis using moving windows aims to test this hypothesis.

5.3.1.3. Moving windows. Fig. 3 shows an example of the dynamic nature of synchrony within the first conversation of pair 2. Along with the raw values of the two speakers (solid green and blue lines), the graphs show the evolution of Pearson's ρ values (dashed red line) for each parameter. The threshold for synchrony/ asynchrony were chosen based on significance tests, ρ values of ± 0.5 indicate significant positive/negative correlations. Significant correlation in this means that the ρ value for the two observed series of numbers is significantly different from 0. The significance level was set at 0.05, indicating that the probability of getting a correlation as large as the observed value by random chance is less than 0.05. Phases of synchrony (highlighted with a red background) and phases of asynchrony (highlighted with a green background) are visible through all parameters. For example, a strong phase of synchrony (denoted by 1) shows parallel patterns as illustrated in Fig. 2; a strong phase of asynchrony (indicated by 2) shows asynchronous behaviour. It is worth noting that the parameters display synchrony and asynchrony in different regions. However, as seen in the region denoted by 3, several parameters may align in synchrony over a certain amount of time.

Table 2 (as well as the Tables A.1–A.3 in Appendix) shows the average number and duration of synchrony and asynchrony phases over the conversations for each of the prosodic features. It is seen, as in Fig. 3, that medianf_0 and sdf_0 exhibit higher number of synchrony phases than sdInt and syllsec. This is further confirmed by the results of an ANOVA analysis with the use of the Tukey–Kramer method to correct for multiple comparisons. Significant differences are reported in the following and marked with * for $p < .05$ and ** for $p < .01$. For all pairs the average number of synchrony and asynchrony phases of sdf_0 and medianf_0 is higher than those of syllsec (all **; except medianf_0 for pair 1 *). With the exception of pair 1, the average number of synchrony and asynchrony phases of sdf_0 and medianf_0 are smaller than those of sdInt (all **; except medianf_0 pair 4 *). Moreover, the average number of synchrony and asynchrony phases of sdf_0 and medianf_0 are significantly different to those of medianInt for pairs 3 and 4 (all **; except medianf_0 pair 4 *). The average number of synchrony (*) and asynchrony (**) phases for medianf_0 are significantly different to those of medianInt for pair 1. In all significant cases the medianf_0 and sdf_0 exhibit the highest average number of synchrony phases as well as the lowest average number of asynchrony phases; syllsec exhibits the lowest average number of synchrony phases and the highest average number of asynchrony phases.

Regarding the average duration of synchrony and asynchrony phases, we found few statistically significant differences but no clear tendency can be drawn from the results across pairs.

Summary. The results of our moving window analysis above reveal that some features show higher levels of synchrony than others (i.e. in particular medianf_0 and sdf_0). This in turn confirms that synchrony dynamically evolves over phases of conversations rather than increases/ decreases continuously over the course of a conversation.

5.3.2. Synchrony dynamics across conversations

Fig. 4 illustrates the dynamic nature of synchrony and asynchrony across conversations for pair 2. The different bar plots represent the evolution of the ratio (or percentage) of synchrony (red bars), asynchrony (green bars) and maintenance (blue bars) obtained for each parameter. As it can be seen in this figure, there is no obvious increase of synchrony and asynchrony across these one week interval conversations for all parameters. These findings are similar for pairs 1, 3 and 4, as the bar plots for these pairs (i.e. Figs. A.1, A.2 and A.3 in Appendix) illustrate.

Summary. These results therefore suggest that prosodic synchrony manifests itself in a dynamic trend within and across conversations.

6. Experiment 2: evaluation of the PAD tool

In a second experiment, we evaluated the relevance of the PAD tool. We focused on two methodological aspects: (1) the HYBRID method and (2) the captured dynamic manifestation of prosodic synchrony.

6.1. Investigating the relevance of the HYBRID method

6.1.1. Objectives and method

In this experiment, we aimed to evaluate the benefits of the HYBRID method compared to the TAMA method. To do so, we have performed independent t-tests ($p < .01$) and investigate whether the extension of the window size for the hybrid method has an effect on the extracted prosodic values, that is, whether prosodic values extracted from TAMA and HYBRID windows are significantly different. We used a window of 20 s-length and a step analysis of 10 seconds for both methods (with the windows being extended to the utterance boundary for the HYBRID method).

6.1.2. Results

Our results show that medianf_0, sdf_0, sdInt and syllsec values are not significantly different for the TAMA and HYBRID methods. Only medianInt is significantly different ($p = 0.0075$). This could be explained by the fact that the window size is quite large and that the extension to the utterance (an not the turn) is quite small. As previously

Table 2
Synchrony and asynchrony phase summary for **Pair 2**. Average number (**AvgN**) of synchrony and asynchrony phases are reported along with their standard deviations (**StdN**). Further, the average duration in number of analysis frames of these phases (**AvgD**) and the standard deviation (**StdD**) over the ten conversations are summarized for each feature.

Feature	Synchrony				Asynchrony			
	AvgN	**StdN**	**AvgD**	**StdD**	**AvgN**	**StdN**	**AvgD**	**StdD**
sdf_0	11.10	3.32	5.55	2.29	6.05	2.89	4.50	3.03
medianf_0	12.40	3.44	4.40	1.71	4.20	1.55	3.40	0.97
medianInt	10.00	1.94	4.80	1.87	6.50	2.32	6.00	1.94
sdInt	5.90	2.18	5.20	2.78	10.20	2.39	5.40	2.22
syllsec	5.20	3.43	4.00	2.05	10.40	3.20	7.60	3.10

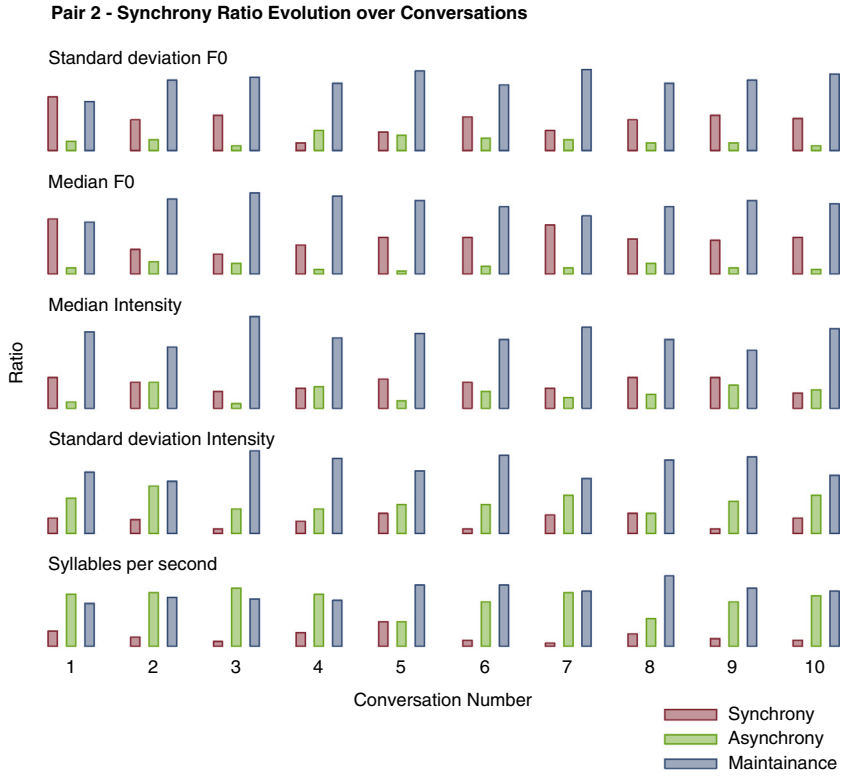

Fig. 4. Evolution of the ratio (i.e. percentage) of synchrony (red), asynchrony (green) and maintenance (blue) for all ten conversations (abscissa) of **Pair 2** for each prosodic parameter (ordinate). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mentioned, window lengths after extension range from 20 s to 26 s. Mean window length is equal to 20.60 s with a standard deviation of 0.68.

Summary. In this context, the HYBRID method does not give different results than the TAMA method.

6.2. Investigating the relevance of a dynamic model

6.2.1. Objectives and method

In order to investigate whether the phases of prosodic synchrony found in the conversations were meaningful and were not capturing accidental or coincidental phenomena, we compared the real synchrony coefficients with the fake ones, using a score transformation method and Mann–Whitney U tests. For this experiment, approximately 2,996 pseudo-conversations were created out of the 41 real conversations and these were analysed using the TAMA model.

We first tested the distribution of each of the five variables to determine whether they were normally distributed for each set. The results indicated that the distribution of each set was non-parametric. Therefore we used a Mann–Whitney U test to determine if the means of the two sets, fake and real, were significantly different. The basic assumption of a Mann–Whitney U test is that the distributions are similarly shaped, which was not the case with the two sets of data. We transformed the data so that the distributions would be similar in order to meet the basic assumption of the Mann–Whitney U. We did this by carrying out a reflect and inverse transformation to ensure that

the two sets of data for each variable were of a similar distribution. The formal for the transformation is given by:

$$\frac{1}{(1+x)-y} \tag{3}$$

where x is the greatest value in the set and y is the actual value of the variable being transformed.

Moreover, we considered the pseudo/fake conversations to be background 'noise' in terms of prosodic accommodation, and therefore the real conversations should be distinguishable from this noise: the real conversations should contain greater instances of strong prosodic accommodation (large positive or negative correlation values). We therefore carried out a z-score transformation of the data for the whole set of conversations (fake and real). We then discriminated based on set membership (fake or real), charting the mean value of each of the 32 analysis windows for each set, in order to determine if there was a distinguishable difference between the two sets of conversations.

6.2.2. Results

Our results indicate that there was a statistically significant difference between the two. The results of the Mann–Whitney U tests indicate that the null-hypothesis be rejected as there is a significant difference between the two. Fig. 5 gives the results of the Mann–Whitney U tests.

The results of the z-score transformation analysis indicate that the fake conversations tended to the mean, with some slight variation, while the real conversations were more distant from the mean. Graphing them out showed a clear separation between the two sets of data. While the distance from the mean (and thus the fake conversational means for each window) varies for, and within, each variable, the pertinent fact is that there is a clear separation between the two sets of conversations. Fig. 6 illustrates the visual differentiation between the fake and real conversations. Moreover Fig. 7 gives details of the mean distances for each group, with the real set demonstrating a clear distance from the fake in most cases. The pitch and pitch range are positively differentiated above the mean while the other variables are negatively differentiated.

Summary. These results therefore support our assumption that the prosodic accommodation within the analysed conversations is not accidental or background accommodation 'noise' but is a product of and an inherent part of the verbal communication between interlocutors.

7. Experiment 3: functional aspects of prosodic synchrony

7.1. Objectives and hypotheses

Prosodic accommodation has been reported to be strongly correlated to the success of information exchange, to the perception of speakers' traits and relation in the interaction, as well as to speakers' engagement in the conversation. In this third experiment, we investigated whether this holds true for our data, and in particular whether these functional aspects can be captured in the detected dynamics of prosodic synchrony.

7.2. Method

To do so, some parts of the conversations were selected and organised in 3 groups according to their levels of synchrony/asynchrony: High (or synchrony, $\rho > .5$), Mid (or maintenance, $\rho \in [-.1, .1]$) and Low (or asynchrony, $\rho < -.5$). While many aligned phases of synchrony in pitch and intensity could be found in our data, only few aligned in pitch and articulation rate or aligned in intensity and articulation rate existed. High, Mid and Low groups for articulation rate synchrony were therefore defined separately from those for pitch and intensity synchrony. Two sets of data were created. The first dataset contains High, Mid and Low Groups of synchrony in pitch (sdf_0 and medianf_0) and intensity (medianInt); the second dataset High, Mid and Low Groups of synchrony in articulation rate (syllsec). These conversation parts were automatically extracted from the first four conversations of each pair. These extracted parts of 110 s-length were then divided into chunks of around 20 s to be used for the perceptual experiment. In total, 195 twenty-second chunks were used for each set.

The perceptual experiment consisted in listening to these chunks and annotating them. Using a 4-point likert scale (−2 strongly disagree; −1 disagree; 1 agree; 2 strongly agree), participants had to answer, for each chunk, 7 statements (a subset and modified version of the questionnaire by Levitan et al., 2011b):

1. The conversation flows naturally.
2. The participants have trouble understanding each other.
3. The speakers are interrupting each other a lot.
4. One of the speakers dominates the conversation. The dominant speaker is talking a lot, is not giving the other the chance to speak.
5. The speakers are involved in the conversation, i.e. they sound interested or engaged in taking part of the conversation.
6. One of the speakers shows more involvement than the other.
7. The speakers seem to like each other.

Participants were recruited using Amazon Mechanical Turk. In total, 114 subjects participated in the experiment. 52.74% were female participants, 47.26 % male. 7.65% were under 20 years old, 62.77% aged between 21 and 40 and 29.58% over 40. A third of the speakers claimed to be fluent in Japanese.

For each subject, we obtained on average 38 annotations (note that 20 subjects annotated more than 100 samples). In total, for dataset 1, 1340 annotations were obtained for synchrony segments (referred to as Group

Test Statistics[a]

	medf0
Mann-Whitney U	44073591.5
Wilcoxon W	4.541E+9
Z	-18.124
Asymp. Sig. (2-tailed)	.000

Test Statistics[a]

	sdf0
Mann-Whitney U	45493319.5
Wilcoxon W	4.081E+9
Z	-11.451
Asymp. Sig. (2-tailed)	.000

Test Statistics[a]

	medianInt
Mann-Whitney U	55935501.0
Wilcoxon W	56796829.0
Z	-6.331
Asymp. Sig. (2-tailed)	.000

Test Statistics[a]

	sdInt
Mann-Whitney U	49426794.5
Wilcoxon W	50288122.5
Z	-12.845
Asymp. Sig. (2-tailed)	.000

Test Statistics[a]

	syllsec
Mann-Whitney U	36361091.0
Wilcoxon W	37219796.0
Z	-25.783
Asymp. Sig. (2-tailed)	.000

Fig. 5. Results of the Mann–Whitney U tests indicating that there is a significant difference between the set of real and pseudo/fake conversations: $p < .05$.

high); 1398 for maintenance segments (referred to as Group mid); and 842 for asynchrony segments (referred to as Group low). Annotations from annotators fluent in Japanese only were distributed as follows: 447 for the Group high; 446 for the Group mid; 284 for the Group low. For dataset 2, 48 annotations were obtained for the Group high; 483 for the Group mid; and 967 for the Group low. Annotations from annotators fluent in Japanese only were distributed as follows: 23 for the Group high; 216 for the Group mid; 419 for the Group low.

The questionnaire time was on average of 71-s long with a std. of 69.20 (if the length was below the length of the segment, then the annotation was rejected). Note that 836 annotations were however rejected (i.e. work of 5 annotators that were obviously cheating).

Anova analyses with Tukey–Kramer correction for multiple testing were performed on the two datasets to compare whether Group low, mid and high were significantly different for the seven annotated statements. In this experiment, the significance level was set at $p < .01$.

7.3. Results

7.3.1. All subjects

7.3.1.1. DataSet 1 (pitch/intensity). Results show that Group low is significantly different from Group mid for statements (4), (5), (6) and (7) and from Group high for statements (1), (4), (5), (6) and (7). However, Group low is not significantly different from Group mid for statements (1), (2) and (3) and from Group high for statements (2) and (3). Also, Group mid is not significantly different from Group high for all statements. Table 3 details the mean and standard deviation values obtained for each group of dataset 1.

7.3.1.2. DataSet 2 (articulation rate). Results show that Group low, mid and high are not significantly different for all statements.

7.3.1.3. Summary. These results suggest that the higher synchrony in pitch and intensity, the more the conversation is perceived as flowing naturally (statement 1) and the more the speakers are perceived engaged (statement 5) and liking each other (statement 7). They also reveal that asynchrony in pitch and intensity indicates that the participation in the conversation is "unbalanced" between the speakers, i.e. one of the speakers shows more engagement (statement 6), dominates or monopolizes more the conversation than the other (statement 4). Synchrony in pitch and intensity however is not correlated to the level of mutual understanding (statement 2) or number of interruptions (statement 3). In addition, synchrony in articulation rate is not correlated to any of the functions investigated in this experiment.

Note that third of these annotations were made by subjects that were fluent in Japanese, the two-third by subjects that were not. T-tests were therefore also performed to investigate whether annotations by speakers who are fluent in Japanese (SFJ) are different from those of speakers who do not speak Japanese (SNFJ). The significance level was set at $p < .01$.

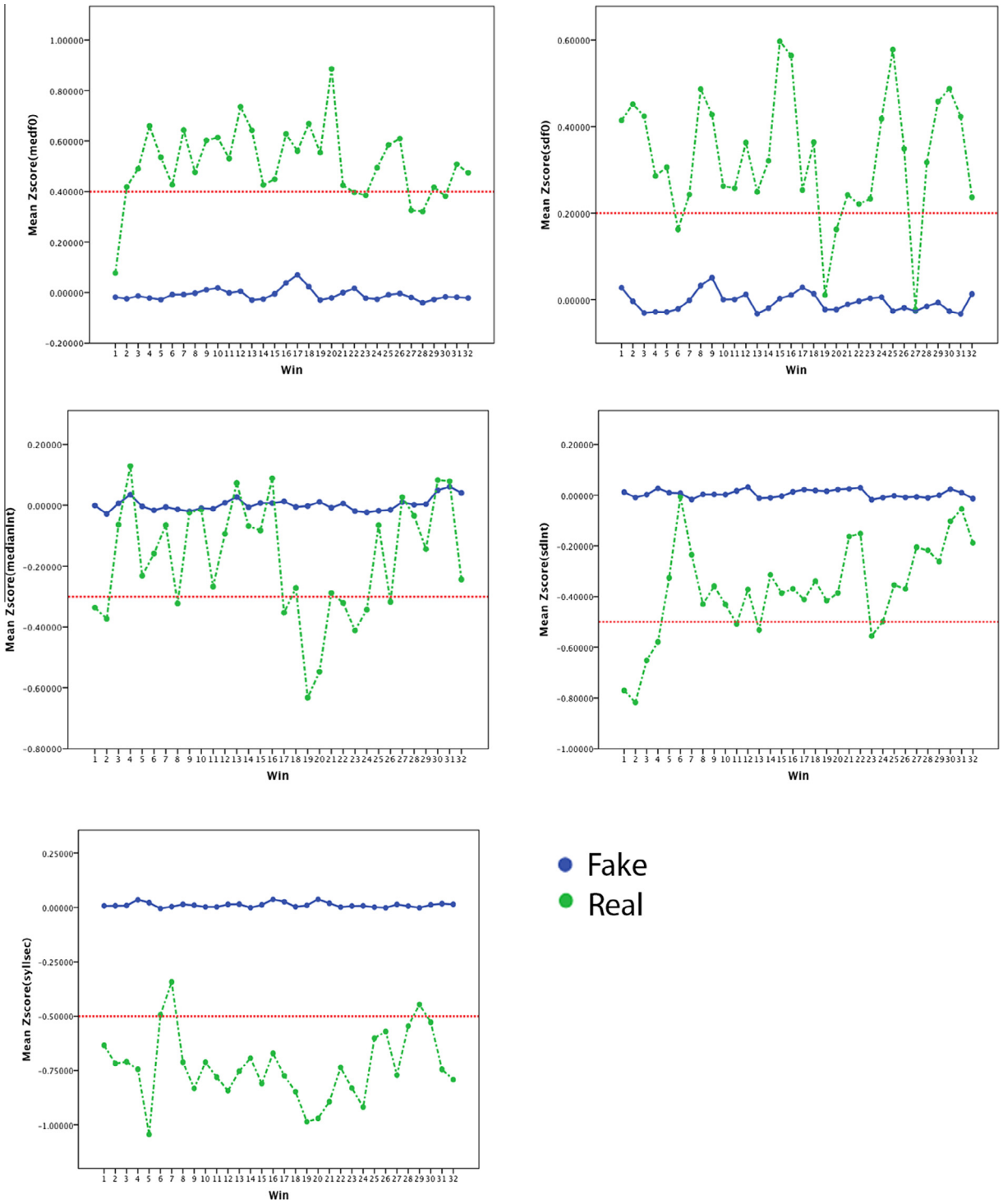

Fig. 6. Results of the z-score transformations for each of the five variables. In each case the real scores are differentiated from the fake scores, which tend to the mean (with some slight variation). The real scores lie at a distance, varying for each variable, from the fake scores. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

7.3.2. Japanese subjects only

7.3.2.1. DataSet 1 (pitch/ intensity). Results show that annotations for Group low by SFJ are not significantly different from those by SNFJ for statements (1), (4), (6) and (7); annotations Group mid by SFJ are not significantly different from those by SNFJ for statements (1), (3), (4), (5) and (6); annotations for Group high by SFJ are not significantly different from those by SNFJ for statements (1), (3), (4), (5) and (7). However, annotations by SFJ differ from those by SNFJ for Group low, statements (2), (3) and (5); for group mid, statements (2) and (7); and for Group high, statements (2) and (6). Table 4 details the means and standard deviation values obtained for each group.

Table 3
Mean (μ) and standard-deviation (sd) obtained for each group (G. low, mid and high) of dataset 1 and related p-values indications (i.e. * for $p < .05$ and for ** for $p < .01$) for group differences (where l-m stands for groups low and mid comparisons, l-h for groups low and high comparisons and m-h for groups mid and high comparisons).

Statements	G. low (μ, sd)	G. mid (μ, sd)	G. high (μ, sd)	p l-m	p l-h	p m-h
1. Conversation flows	1.22, 0.89	1.24, 0.85	1.31, 0.82		**	
2. Mutual understanding	−1.00, 1.10	−1.09, 1.04	−1.12, 1.04			
3. Interruptions	−0.83, 1.13	−0.72, 1.19	−0.72, 1.21			
4. Holding the floor	0.48, 1.39	0.12, 1.37	0.11, 1.40	**	**	
5. Speakers' overall engagement	0.81, 1.10	0.95, 1.07	1.03, 1.06	**	**	
6. Speakers' individual engagement	0.73, 1.27	0.42, 1.32	0.31, 1.33	**	**	
7. Affinity	0.75, 1.10	0.97, 1.02	1.06, 1.00	**	**	

7.3.2.2. DataSet 2 (articulation rate). The results show that SNFJ significantly rated the flow (i.e. statement 1) of the conversation lower for condition low, with SFJ ($\mu = 1.39$ and $\sigma = 0.96$) and SNFJ ($\mu = 1.25$ and $\sigma = 0.94$) and $p = 0.02$. Additionally, statement (2) was rated significantly lower for conditions low and mid by SNFJ, with SFJ low ($\mu = -1.01$ and $\sigma = 1.35$); SFJ mid ($\mu = -0.94$ and $\sigma = 1.35$); SNFJ low ($\mu = -1.17$ and $\sigma = 0.88$); and SNFJ mid ($\mu = -1.19$ and $\sigma = 0.81$). All other statements did not reveal any significant differences between groups SFJ and SNFJ.

7.3.2.3. Summary. Overall, for dataset 1, annotations by SFJ are not significantly different from those by SNFJ except for statement (2), where it is different for groups low, mid and high. Similarly to the differences between groups SFJ and SNFJ for dataset 1, significant differences could be identified for dataset 2, where phases of synchrony in articulation rate are investigated. Here, SNFJ rate the perceived understanding significantly lower than SFJ for conditions low and mid. Additionally, for statement 1 SFJ rate the flow higher than SNFJ for the condition low.

T-tests were then performed to investigate the correlation between the perceived level of speakers' mutual understanding and the level of synchrony in pitch and intensity for SFJ only. Results show that Group low ($\mu = -0.68$, $sd = 1.36$) is not significantly different from Group mid ($\mu = -0.85$, $sd = 1.27$), Group low is not significantly different from Group high ($\mu = -0.78$, $sd = 1.31$) and Group mid is not significantly different from Group high.

7.3.3. Overview

These results therefore suggest that the dynamics in pitch and intensity synchrony inform about the naturalness of the conversation flows, speakers' engagement as well as their perceived affinity. Overall, the functional role of prosodic synchrony is not perceived differently between native and non-native speakers.

8. Discussion

In this paper, we have presented a tool for the automatic measurement of prosodic accommodation dynamics in social interaction. The tool was built under the assumption that prosodic accommodation does not only continuously increase over the course of a conversation or across several conversations but rather varies dynamically, parallel to the dynamic nature of social interaction. Only a few studies have examined the dynamic manifestation of prosodic synchrony and we hypothesized that prosodic accommodation undergoes dynamic changes over the course of an interaction, as interlocutors do not remain involved to the same degree over the whole course of a conversation. The dynamic manifestation has a social-functional role to play, being related to the situational context.

8.1. Measuring the dynamic manifestation of prosodic synchrony

Due to the fact that prosodic accommodation has been shown to increase continuously as well as to vary over the course of the conversation, we assumed that it can actually

Table 4
Mean (μ) and standard-deviation (sd) obtained for each group (G. low, mid and high) and for each group of annotators (where SFJ stands for Speakers Fluent in Japanese and SNFJ Non-Fluent in Japanese), with related p-values indications for SFJ-SNFJ group differences (i.e. * for $p < .05$ and ** for $p < .01$).

Sts	SFJ-l μ, sd	SNFJ-l μ, sd	SFJ-m μ, sd	SNFJ-m μ, sd	SFJ-h μ, sd	SNFJ-h μ, sd	G. low p	G. mid p	G. high p
1.	1.27, 0.84	1.20, 0.91	1.30, 0.82	1.22, 0.86	1.30, 0.86	1.32, 0.80			
2.	−0.68, 1.36	−1.17, 0.90	−0.85, 1.27	−1.20, 0.89	−0.78, 1.31	−1.29, 0.82	**	**	**
3.	−0.66, 1.29	−0.92, 1.04	−0.59, 1.38	−0.78, 1.09	−0.68, 1.31	−0.74, 1.15	**		
4.	0.47, 1.36	0.49, 1.40	0.29, 1.32	0.05, 1.39	0.36, 1.32	−0.02, 1.42			**
5.	0.95, 1.09	0.74, 1.10	1.02, 1.11	0.92, 1.05	1.07, 1.06	1.02, 1.05	*		
6.	0.60, 1.31	0.8, 1.25	0.40, 1.31	0.43, 1.32	0.40, 1.34	0.26, 1.33			*
7.	0.77, 1.16	0.73, 1.06	0.84, 1.19	1.03, 0.93	0.99, 1.08	1.10, 0.95		**	

manifest as both a continuous and dynamic phenomenon. Some cues may exhibit a largely linear rate of accommodation, increasing or decreasing in a continuous fashion over the course of an interaction or across several consecutive interactions. Likewise, accommodation between some parameters can fluctuate over the course of a conversation or across several conversations.

We hypothesised that the continuous and dynamic manifestation may depend on a number of factors. In a form-function mapping, we suggested that when the functional aspect of accommodation is static over the course of a conversation, it is manifested in a continuous trend, while when if it is linked to speakers social states and intentions, it is exhibited in a dynamic trend, where accommodation fluctuates with the conversational partners' changes of states. One may for instance observe a dynamic manifestation of accommodation as it fluctuates with the speakers' degree of involvement in a conversation (De Looze et al., 2011; De Looze and Rauzy, 2011).

In this study, we first investigated the continuous manifestation of synchrony by examining the amount of synchrony at the start and end of the interactions. We also segmented the conversations into three parts and examined the amount of synchrony in each of the three segments. The results indicated that there was no significant difference between the first half and second half of all the conversations, and that there was no significant difference between the first, second or third segments. We further analysed the evolution of prosodic synchrony at different time intervals and found that prosodic synchrony varies several times over the course of all conversations. Our results therefore corroborate previous results reported for English in De Looze et al. (2011), De Looze and Rauzy (2011) and Vaughan (2011).

This supports the original hypothesis that prosodic synchrony is a dynamic phenomenon and that the continuous increase fails to take this into account; were we to rely solely on this measure, we would have missed important communicative aspects of prosodic accommodation, as evidenced by the results of the annotation procedure. The length of the conversations (approximately 30 min), most likely influenced the result of the linear analysis. A simple segmentation of the speech into halves or thirds is not adequate for speech of this length and is more likely suited to the analysis of shorter speech segments. In this instance, the PAD tool is more suited: a 30 min conversation has more scope to cover a diverse range of topics; better facilitates the emergence of interpersonal communicative strategies and greater scope for differing amounts of conversational flow and involvement.

It is worth noting too that the different prosodic parameters, pitch, intensity and articulation rate, do not exhibit the same amount of synchrony in the analysed conversations. We have shown that in all pairs and all conversations, medianf_0 and sdf_0 exhibit the highest average number of synchrony phases as well as the lowest average number of asynchrony phases. On the contrary, syllsec exhibits the lowest average number of synchrony phases and the highest average number of asynchrony phases. MedianInt and sdInt also exhibit high levels of asynchrony phases.

Pitch is one of the most important prosodic parameters in verbal communication. It has been found to be strongly correlated with the activation dimension of emotional models (Juslin and Scherer, 2005; Banse and Scherer, 1996) and the active aspect of emotional categories. Research has suggested that all languages use variations in pitch to signal meaning and affect (Scherer and Wallbott, 1994). If, as already mentioned in the introduction, one considers a conversation to be an interactive dialogue where interlocutors do not remain involved to the same degree over the whole course of a conversation, it can be assumed that speaker's states may change over the course of the interaction and that they may be mainly expressed and recognized by changes in the amount of pitch synchrony. This has been shown for instance in De Looze and Rauzy (2011) and De Looze et al. (2011) and in the present study where speakers' pitch synchrony is reported to be highly correlated to speakers' degree of involvement. The perception of the dynamic aspect of a speaker's states in terms of pitch would be all the more prominent as the human auditory system is very sensitive to changes in pitch, with the audible frequency range of the human auditory system ranging from 20 Hz to 20 kHz (Rumsey and McCormick, 2002); the consequence of this being that speakers show pitch adaptation much more than other types of prosodic adaptation. This would be particularly likely, for instance, in male-female pairs, as converging in pitch would require too much vocal effort[1] and would affect a speaker's vocal identity too much (as specified for instance in Ohala's frequency code (Ohala, 1983)).

On the contrary, small changes in articulation rate may not be as well perceived as small changes in pitch. If one considers the strong link between perception and production, synchronising variations in articulation rate may therefore not be intended. Furthermore, studies have shown that variations in speech rate are rather due to variations in the number of pauses and their mean duration than to variations in the actual articulation rate (Goldman-Eisler, 1968; Grosjean and Deschamps, 1975). This is exemplified in our previous study where we found high degrees of synchrony in terms of pause duration (De Looze and Rauzy, 2011; De Looze et al., 2011) and in our present study for which we observed very few levels of synchrony in articulation rate. As commented by Goldman-Eisler (1961), articulation rate is "*a personality constant of remarkable invariance*". If speakers' articulation rate is rather constant in nature, one may therefore adapt their speech rate in terms of pause number and duration rather than in terms of articulation rate.

[1] The average male f_0 range is approximately 75–300 Hz, the average female f_0 range 100–500 Hz.

Regarding asynchrony phases in intensity, one could assume that the nature of conversation naturally implies such a pattern, as while one speaker is talking, the interlocutor is remaining partly silent or back-channeling. This would suggest that synchrony phases, on the contrary, are rather indicative of overlapped speech.

8.2. Prosodic synchrony over conversations

In this study, we also investigated the dynamic nature of prosodic synchrony across conversations. Our data had the advantage that the participants did not know each other prior to taking part in the corpus recordings. This allowed us to examine whether the amount of synchrony increased as the participants became better acquainted.

We hypothesized that over the course of several conversations, even if participants would get better acquainted, because of the dynamic nature of social interaction, prosodic synchrony would not continuously increase across conversations but be specific to the interaction (following the premises of Interactional Linguistics (Mondada, 2001)).

Our results support our assumption. Prosodic accommodation varies over each conversation, suggesting that the amount of synchrony is more than likely affected by other factors such as speakers' degree of involvement, their role in the conversation (e.g. dominant vs. dominated) as well as the adoption of interpersonal communication strategies. Considering our original hypothesis that prosodic synchrony is a dynamic phenomenon, this should be expected; regardless of acquaintance, the dynamic manifestation is related to social-functional phenomena. As seen in the literature review at the start of the paper, numerous researchers have found prosodic accommodation to be related to a number of important social phenomena – likeability, common-ground, cognitive alignment, engagement – and to be used for many different purposes, e.g. as a vocal manifestation of social-hierarchical distance, a method of achieving social approval and acceptance (see Section 2.3).

8.3. Functional role of prosodic synchrony

In this study, the functional role of prosodic synchrony dynamics was further investigated. In particular, we have investigated seven possible functional aspects: how they inform about (1) the perceived naturalness of the conversation flow, (2) speakers' mutual understanding, (3) speech interruptions, (4) aspect of dominance (in terms of holding the floor), (5) speakers' degree of overall and individual engagement and (6) speaker's affinity (in terms of positive perception of the interaction).

Our results have first shown that the higher synchrony in pitch and intensity, the more the conversation is perceived as flowing naturally and the more the speakers are perceived engaged and liking each other. They also reveal that an asynchrony in pitch and intensity indicates that the participation in the conversation is 'unbalanced' between the speakers, i.e. one of the speakers shows more engagement, dominates or monopolizes more the conversation than the other.

These results confirm earlier studies which revealed the importance of prosodic accommodation in facilitating the exchange of information and in increasing the social success of the interaction. The fact that when speakers' degree of prosodic synchrony is high, the conversation is perceived as flowing more naturally, the speakers are perceived as more engaged and liking each other, and the fact that, on the contrary, asynchrony is associated with 'unbalanced' behaviours or attitudes, suggest that prosodic synchrony is a strong indicator of speakers' social resonance in the conversation. Maintenance and differentiation are, on the contrary, used by speakers to maintain social distance with their interlocutors (Giles et al., 1991). These results instantiate Chartrand and Bargh's comment on the chameleon effect (Chartrand and Bargh, 1999) – it operates to create greater liking and to ease the interaction.

Our study further corroborates earlier findings we observed in natural conversations of English (De Looze and Rauzy, 2011; De Looze et al., 2011) on the link between prosodic synchrony and speakers' degree of involvement. It is interesting to note that for two different languages, and in particular different cultures, a similar process seem to be used to convey similar functions. The fact that no clear difference was found in the perception of native and non-native speakers of Japanese further suggests that prosodic accommodation has become automatic over the course of human evolution (Chartrand and Bargh, 1999) and has been used to serve specific functions in social interaction. It would be worth performing systematic cross-languages comparisons to better understand the universal vs. language-specific nature of this process and its functions.

Furthermore, we have not observed any correlation between synchrony in pitch and intensity and the perceived level of mutual understanding, the annotations being performed either by native or non-native speakers of Japanese. Nor have we observed any link between synchrony in pitch and intensity and the number of perceived interruptions.

Accommodation is performed at many different domains (semantic, linguistic, syntactic, prosodic) and according to different modalities (verbal, vocal, visual). One possible explanation is that mutual understanding is enhanced through the alignment of other domains such as the syntactic and lexical. Prosodic accommodation, in the vicinity of backchannels, has yet been found to augment conversational partners' performance on their joint task (Levitan et al., 2011a; Levitan and Hirschberg, 2011). It may therefore also serves information exchange, at a more local level. It would be worth investigating how accommodation, at different domains and modalities, and at different temporal spans, contributes to facilitate information exchange as well as to augment the social success of the interaction.

Finally, our results reveal that synchrony in articulation rate is not correlated to any of the functional aspects investigated in this study. This however does not exclude the possibility that synchrony in articulation rate may convey other functions in social interactions.

8.4. Relevance of the PAD tool

In this study, we have also evaluated the relevance of our tool for the automatic measurement of prosodic accommodation. The evaluation concerned the use of the HYBRID method to extract prosodic cues as well as the capture of prosodic accommodation dynamics.

8.4.1. The HYBRID method

Current approaches to extract prosodic cues comprise two types of methods: the utterance-based (Levitan and Hirschberg, 2011) and the TAMA methods (Kousidis et al., 2008). We have proposed in our tool to use a HYBRID method, based on both. The argument for such a method is that it allows for the extraction of prosodic features at different temporal spans and therefore for the investigation of prosodic accommodation at different levels of the interaction. Such a method, sensitive to utterance boundaries, also enables the extraction of prosodic features from entire utterances, which are representative of the utterance prosody and its functions in the interaction. We evaluated in this study the benefits of the HYBRID method compared to the TAMA method.

Our results do not support the prevalence of the HYBRID method over the TAMA method. As already mentioned, this could be due to the fact that, in this experiment, the extension of the HYBRID moving windows is not important enough compared to the windows default length to result in large differences in prosodic features values when compared to the TAMA extracted values. It can be assumed, however, that these values would be different and the HYBRID method necessary when measuring accommodation in terms of intonation patterns as intonation patterns span and functionally make sense at the utterance level. Further analysis is needed to compare the performance of both methods, as well as the utterance-based method, testing different window sizes and different steps of analysis and relating them to the functional aspects of prosodic accommodation.

In this experiment, we have not investigated the dynamics of prosodic accommodation on a sub-utterance level as was done by Heldner et al. (2010) Levitan et al. (2011a) and Levitan and Hirschberg (2011) even though we believe that this phenomenon can be more local that what we have investigated in our study. As we explained, we believe that prosodic accommodation spans over different temporal domains, which justifies an investigation at a sub-utterance level, at the utterance level as well as on larger domains (several utterances).

In Levitan et al, prosodic accommodation was reported in the vicinity of feedbacks (over windows of 200 to 1000 ms length). This short-term prosodic adaptation was found to be strongly correlated to the speakers conversational coordination and performance on their joint task; the higher prosodic adaptation, the better the coordination between partners, the better they perform their task.

In our study, prosodic features were extracted every 10 s (over windows of 20 s length with an overlap of 10 s). We have previously mentioned that this longer-term prosodic accommodation is strongly correlated to speakers' degree of involvement, to the perceived naturalness of the conversation flow as well as to speakers social closeness or rapport.

Both methods reveal different aspects of prosodic accommodation: large windows would capture 'social closeness', while shorter windows would inform about the information exchange process.

8.4.2. Prosodic accommodation dynamics measurement

Compared to current methods which examine prosodic accommodation as a linear phenomenon only and which measure its amount at the beginning and end of conversations, our tool also allows for its measurement at different time intervals, quantifying its degree using a set of overlapping moving windows. In our study, we evaluated whether the captured dynamics were not accidental phenomena but meaningful in the frame of social interaction.

Similarly to Ramseyer and Tschacher (2010), we created a number of pseudo-conversations for which we computed synchrony coefficients. These coefficients were then compared to those obtained with real conversations using a Mann–Whitney U test and by carrying out a z-score transformation on the complete set of data (real and pseudo)

Table A.1
Synchrony and asynchrony phase summary for **Pair 1**. Average number (**AvgN**) of synchrony and asynchrony phases are reported along with their standard deviations (**StdN**). Further, the average duration in number of analysis frames of these phases (**AvgD**) and the standard deviation (**StdD**) over the ten conversations are summarized for each feature.

Feature	Synchrony				Asynchrony			
	AvgN	**StdN**	**AvgD**	**StdD**	**AvgN**	**StdN**	**AvgD**	**StdD**
sdf_0	8.65	3.18	5.35	2.08	5.55	2.23	4.10	1.60
medianf_0	10.15	3.00	4.65	1.16	4.50	1.65	5.10	2.77
medianInt	6.30	3.23	5.30	2.63	9.20	2.30	5.50	3.34
sdInt	7.90	2.02	7.20	3.29	6.10	3.03	4.30	2.41
syllsec	4.30	3.13	3.40	2.07	11.00	3.46	4.80	1.87

Table A.2
Synchrony and asynchrony phase summary for **Pair 3**. Average number (**AvgN**) of synchrony and asynchrony phases are reported along with their standard deviations (**StdN**). Further, the average duration in number of analysis frames of these phases (**AvgD**) and the standard deviation (**StdD**) over the ten conversations are summarized for each feature.

Feature	Synchrony				Asynchrony			
	AvgN	**StdN**	**AvgD**	**StdD**	**AvgN**	**StdN**	**AvgD**	**StdD**
sdf_0	10.82	3.40	5.27	2.49	3.55	1.13	3.45	1.69
medianf_0	11.55	2.46	5.09	1.64	2.82	1.83	3.36	3.17
medianInt	4.45	2.91	4.91	1.64	8.91	2.70	5.27	3.00
sdInt	2.18	2.82	1.55	2.11	11.55	2.38	5.36	1.75
syllsec	2.64	1.57	3.09	2.21	11.55	1.97	6.27	1.62

Table A.3
Synchrony and asynchrony phase summary for **Pair 4**. Average number (**AvgN**) of synchrony and asynchrony phases are reported along with their standard deviations (**StdN**). Further, the average duration in number of analysis frames of these phases (**AvgD**) and the standard deviation (**StdD**) over the ten conversations are summarized for each feature.

Feature	Synchrony				Asynchrony			
	AvgN	**StdN**	**AvgD**	**StdD**	**AvgN**	**StdN**	**AvgD**	**StdD**
sdf_0	7.80	1.48	5.50	2.55	6.10	1.97	5.00	3.33
medianf_0	11.90	2.77	4.40	1.26	4.90	1.66	4.00	1.33
medianInt	4.50	2.37	4.10	2.69	11.20	2.44	6.70	2.21
sdInt	4.50	3.14	5.60	3.92	10.90	2.77	5.30	0.95
syllsec	2.10	1.91	2.90	2.13	11.20	1.75	5.20	1.75

and discriminating based on group membership. Our study reveals that prosodic accommodation in real conversations is significantly different to that in pseudo-conversations. We take this as evidence that the captured dynamic manifestation of prosodic accommodation was not a random phenomenon and that the chosen window size (of 100 s-length) enables to capture salient aspects of the phenomenon. The veracity of our tool was further confirmed by our experiment on the functional role of prosodic synchrony dynamics in social interaction for four of the variables examined. The fifth, syllables per second (articulation rate), was not correlated with any functional aspect. Considering the articulation rate in the real conversations was significantly different to that of pseudo-conver-

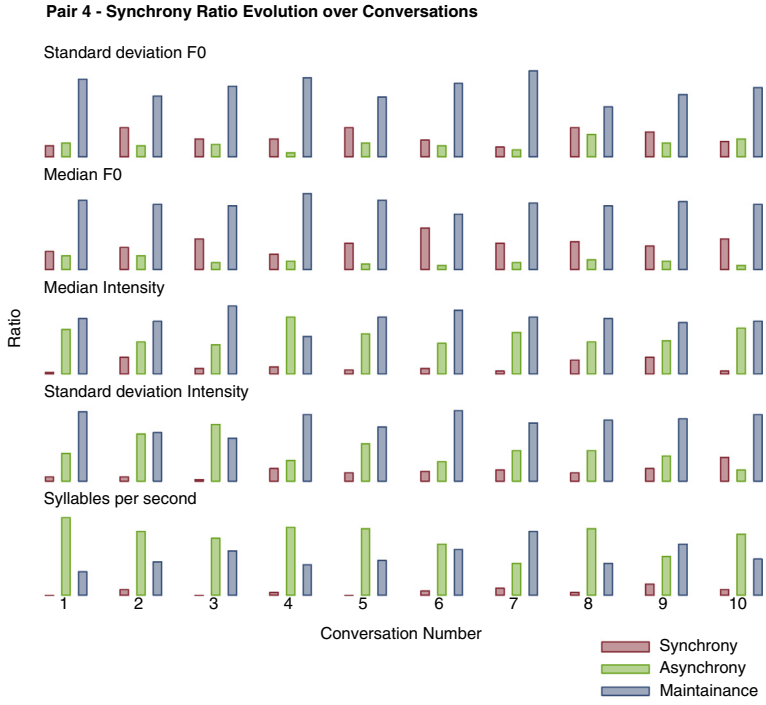

Fig. 7. Mean distances obtained for fake and real conversation groups.

sations, this is more than likely due to the questions asked in the perceptual test not capturing or relating to articulation rate. Future work will need to be carried out to determine the functional aspects of articulation rate.

While a smaller window size may also do the same (and a larger window size), the question of granularity is one that needs to be examined in great detail. It may be that there is no one optimum window size that can be applied in a general sense; rather that the window size depends on the level of granularity and span that one would like to investigate. It is out hypothesis that long stretches of prosodic synchrony between interlocutors is capturing important functional aspects of synchrony. The view that we take is that long stretches of synchrony are as salient as smaller levels and are capturing different, yet meaningful aspects. While a smaller window and sub-utterance level analysis is necessary in examining phenomena such as backchannels, longer durations above the utterance level is necessary to examine the social aspects of synchrony. We believe that there is no optimal temporal span to search for; one should rather look for a temporal span suitable for to the type of functional analyses and features to be investigated.

Median

type	medf0
F	-.008970
R	.252390
Total	-.005668

Median

type	sdf0
F	-.003111
R	.147600
Total	-.000923

Median

type	medianInt
F	-.009453
R	-.088468
Total	-.010579

Median

type	sdInt
F	-.002882
R	-.200725
Total	-.004957

Median

type	syllsec
F	-.000869
R	-.406370
Total	-.006122

Fig. A.1. Evolution of the ratio (i.e. percentage) of synchrony (red), asynchrony (green) and maintenance (blue) for all ten conversations (abscissa) of **Pair 1** for each prosodic parameter (ordinate). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. A.3. Evolution of the ratio (i.e. percentage) of synchrony (red), asynchrony (green) and maintenance (blue) for all ten conversations (abscissa) of **Pair 4** for each prosodic parameter (ordinate). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

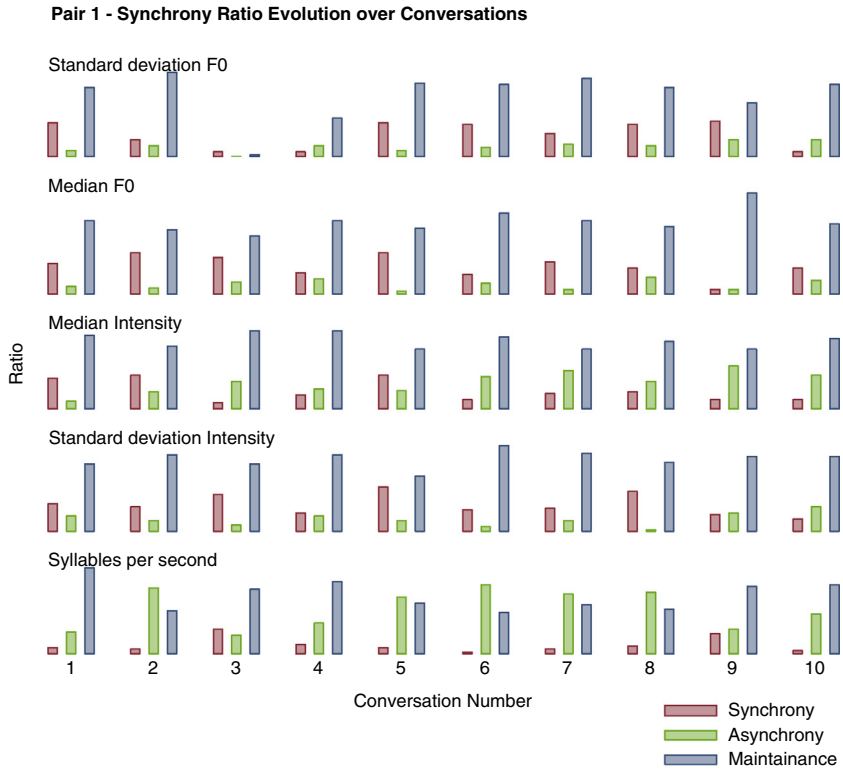

Fig. A.2. Evolution of the ratio (i.e. percentage) of synchrony (red), asynchrony (green) and maintenance (blue) for all ten conversations (abscissa) of **Pair 3** for each prosodic parameter (ordinate). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

9. Conclusions

In this paper we presented a tool for the measurement of the prosodic synchrony that is ideally suited to the capture of its dynamic manifestation. Using this system, we investigated prosodic synchrony in Japanese dyadic telephone conversations in a number of different ways: we examined synchrony as a dynamic phenomenon; we examined the social-communicative functional role of prosodic synchrony using crowd-sourced annotations; we compared our results to pseudo-synchrony measurements derived from randomly chosen artificial conversations and determined that our tool was capturing salient moments of prosodic synchrony through the use of paired t-tests.

We examined prosodic accommodation over individual conversations and across groups of conversations carried out with the same conversational partners, the results of which support our hypothesis that prosodic synchrony is a largely dynamic phenomenon. Overall, our study shows that the degree of prosodic synchrony changes dynamically over the course of a conversation and across conversations, and that these dynamics inform about the naturalness of the conversation flow, the speakers' degree of involvement and their affinity in the conversation.

10. Future work

In the presentation of the PAD tool, we have proposed a set of seven states underlying prosodic accommodation. In this paper, we have focused on its synchronous/asynchronous forms. Future work will therefore focus on the development of a suitable measurement for its convergent and divergent forms. This will enable to investigate whether synchrony and convergence temporally co-occur and convey the same functions in social interaction.

Different window sizes, in the investigation of the evolution of prosodic accommodation, will also be tested in order to better understand its role at both local and global levels. As previously postulated, we believe that prosodic accommodation is displayed at different temporal spans, to convey different functions.

Accommodation has been reported at many other different levels: speech sounds, syntax, lexicon, body movement, gestures and so on. Recent investigations have looked at accommodation from a multimodal perspective and have suggested that convergent features at a certain level may have been mediated by convergent signals at other levels. For future work, we intend to extend our tool by measuring accommodation in terms of voice quality, head movement quantity and body activity.

Acknowledgments

This work was undertaken as part of the FASTNET project – Focus on Action in Social Talk: Network Enabling Technology funded by Science Foundation Ireland (SFI) 09/IN.1/I2631.

Appendix A. Additional results

Tables A.1–A.3 and Figs. A.1–A.3.

References

Agarwal, S.K., Paek, T., Rajput, N., Thies, B., 2011. 2nd international workshop on intelligent user interfaces for developing regions: IUI4DR. IUI 2011.

Apple Inc. Apple SIRI Homepage, 2011.

Aubanel, V., Nguyen, N., 2010. Automatic recognition of regional phonological variation in conversational interaction. Speech Communication 52 (6), 577–586.

Babel, M., Bulatov, D., 2011. The role of fundamental frequency in phonetic accommodation. Language and Speech 55 (2), 231–248.

Bailly, G., Lelong, A., 2010. Speech dominoes and phonetic convergence. In: Interspeech 2010, pp. 1153–1156.

Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70 (3), 614–636.

Bavelas, J.B., Black, A., Lemery, C.R., Mullett, J., 1986. I show how you feel: motor mimicry as a communicative act. Journal of Personality and Social Psychology 50 (2), 322.

Bell, L., Gustafson, J., Heldner, M., 2003. Prosodic adaptation in human–computer interaction. In: Proceedings of ICPhS, vol. 3. Citeseer, pp. 833–836.

Bernieri, F.J., Rosenthal, R., 1991. Interpersonal coordination: behavior matching and interactional synchrony. Fundamentals of Nonverbal Behavior, 401.

Black, J.W., 1949. The intensity of oral responses to stimulus words. Journal of Speech and Hearing Disorders 14 (1), 16.

Boersma, P., Weenink, D., 2006. Praat: Doing Phonetics by Computer. Available from: <www.praat.org>.

Boylan, P., 2004. Accommodation theory revisited. Technical report, University of Rome III (Italy), Rome.

Branigan, H.P., Pickering, M.J., Pearson, J., McLean, J.F., 2010. Linguistic alignment between people and computers. Journal of Pragmatics 42 (9), 2355–2368.

Breazeal, C., 2002. Regulation and entrainment in human–robot interaction. The International Journal of Robotics Research 21 (10–11), 883–902.

Brennan, S.E., 1996. Lexical entrainment in spontaneous dialog. In: Proceedings of ISSD, pp. 41–44.

Burgoon, J.K., Stern, L.A., Dillman, L., 1995. Interpersonal adaptation: dyadic interaction patterns. In: Number Cambridge. Cambridge University Press, UK.

Campbell, N., 2004. Speech and expression; the value of a longitudinal corpus'. In: Proceedings Fourth International Conference on Language Resources and Evaluation (LREC'04), pp. 183–186.

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.

Cleland, A.A., Pickering, M.J., 2003. The use of lexical and syntactic information in language production: evidence from the priming of noun-phrase structure. Journal of Memory and Language 49 (2), 214–230.

Collins, B., 1998. Convergence of fundamental frequencies in conversation: if it happens, does it matter? In: Fifth International Conference on Spoken Language Processing.

Condon, W.S., Sander, L.W., 1974. Synchrony demonstrated between movements of the neonate and adult speech. Child Development, 456–462.

Coulston, R., Oviatt, S., Darves, C., 2002. Amplitude convergence in children's conversational speech with animated personas. Proceedings of the Seventh International Conference on Spoken Language Processing, vol. 4. Citeseer, pp. 2689–2692.

Crowne, D.P., Marlowe, D., 1960. A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology 24 (4), 349.

De Jong, N.H., Wempe, T., 2009. Praat script to detect syllable nuclei and measure speech rate automatically. Behavior Research Methods 41 (2), 385–390.

De Looze, C., 2010. Analyse et interprétation de l'empan temporel des variations prosodiques en français et en anglais contemporain. PhD Thesis, Doctoral Thesis, Université de Provence.

De Looze, C., Rauzy, S., 2011. Measuring speakers' similarity in speech by means of prosodic cues: methods and potential. Proceedings of Interspeech 2 011. ISCA, pp. 1393–1396.

De Looze, C., Oertel, C., Rauzy, S., Campbell, N., 2011. Measuring dynamics of mimicry by means of prosodic cues in conversational speech. Proceedings of ICPhS. Springer, pp. 1294–1297.

Delvaux, V., Soquet, A., 2007. The influence of ambient speech on adult speech productions through unintentional imitation. Phonetica 64 (2–3), 145–173.

Edlund, J., Heldner, M., Hirschberg, J., 2009. Pause and gap length in face-to-face interaction. In 10th Annual Conference of the International Speech Communication Association, pp. 2779–2782.

Ferguson, C.A., 1975. Toward a characterization of english foreigner talk. Anthropological Linguistics 17 (1), 1–14.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., Fukui, I., 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. Journal of Child Language 16 (03), 477–501.

Gallois, C., Callan, V.J., 1988. Communication accommodation and the prototypical speaker: predicting evaluations of status and solidarity. Language and Communication 8 (3), 271–283.

Gallois, C., Callan, V.J., 1991. Interethnic accommodation: the role of norms. Contexts of Accommodation: Developments in Applied Sociolinguistics, 245–269.

Giles, H., Coupland, N., Coupland, J., 1991. Accommodation theory: communication, context, and consequence. Contexts of Accommodation: Developments in Applied Sociolinguistics, 1–68.

Goldman-Eisler, F., 1961. The significance of changes in the rate of articulation. Language and Speech 4 (3), 171–174.

Goldman-Eisler, F., 1968. Psycholinguistics: Experiments in Spontaneous Speech. Academic Press.

Google, 2011. Google Voice Search.

Gregory, S.W., Dagan, K., 1997. Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality. Journal of Nonverbal Behavior 21 (1), 23–43.

Gregory, S.W., Hoyt, B.R., 1982. Conversation partner mutual adaptation as demonstrated by Fourier series analysis. Journal of Psychological Research 11, 35–46.

Gregory, S.W., Webster, S., Huang, G., 1993. Voice pitch and amplitude convergence as a metric of quality in dyadic interviews. Language and Communication 13, 195–217.

Grosjean, F., Deschamps, A., 1975. Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation. Phonetica 31 (3–4), 144–184.

Haywood, S.L., Pickering, M.J., Branigan, H.P., 2005. Do speakers avoid ambiguities during dialogue? Psychological Science 16 (5), 362–366.

Heldner, J., Edlund, M., Hirschberg, J., 2010. Pitch similarity in the vicinity of backchannels. In: Proceedings of Interspeech 2010, pp. 1–4.

Hess, U., Blairy, S., 2001. Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. International Journal of Psychophysiology 40 (2), 129–141.
Jaffe, J., 2001. Rhythms of Dialogue in Infancy: Coordinated Timing in Development. Wiley-Blackwell.
Jaffe, J., Feldstein, S., 1970. Rhythms of Dialogue, Academic Press, New York.
Juslin, P.N., Scherer, K.R., 2005. Vocal expression of affect. In: The New Handbook of Methods in Nonverbal Behavior Research. Oxford University Press, Oxford, UK.
Kleinberger, T., Becker, M., Ras, E., Holzinger, A., 2007. Ambient intelligence in assisted living : enable elderly people to handle future interfaces. Access, 103–112.
Kopp, S., 2010. Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. Speech Communication 52 (6), 587–597.
Kousidis, S., Dorran, D., McDonnell, C., Coyle, E., 2008. Times series analysis of acoustic feature convergence in human dialogues. In: Proceedings of Interspeech.
Kousidis, S., Dorran, D., Mcdonnell, C., Coyle, E., 2009. Convergence in human dialogues time series analysis of acoustic feature. In: Proceedings of SPECOM 2009, St. Petersburg, Russia, p. 2.
Lakin, J.L., Chartrand, T.L., 2003. Using nonconscious behavioral mimicry to create affiliation and rapport. Journal of Psychological Science 14 (4), 334–339.
Lee, C.C., Black, M., Katsamanis, A., Lammert, A.C., Baucom, B.R., Christensen, A., Georgiou, P.G., Narayanan, S.S., 2010. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In 11th Annual Conference of the International Speech Communication Association, pp. 793–796.
Levelt, W.J.M., 1982. Linearization in describing spatial networks. Processes, Beliefs, and Questions, 199–220.
Levitan, R., Hirschberg, J., 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In: 12th Annual Conference of the International Speech Communication Association, 2011.
Levitan, R., Gravano, A., Hirschberg, J., Entrainment in speech preceding backchannels. In: Proc. of ACL 2011, pp. 113–117.
Levitan, R., Gravano, A., Willson, L., 2011. Acoustic-prosodic entrainment and social behavior, In: INTERSPEECH, pp. 3081–3084.
Lu, H., Brush, A., Priyantha, B., Karlson, A.K., Liu, J., 2011. SpeakerSense: energy efficient unobtrusive speaker identification on mobile phones. Work, 188–205.
Maganti, H.K., Motlicek, P., Gatica-Perez, D., 2007. Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), vol. 4. IEEE, pp. IV-1037–IV-1040.
Matarazzo, J.D., Wiens, A.N., 1967. Interviewer influence on durations of interviewee silence. Journal of Experimental Research in Personality 2 (1), 56–69.
Maurer, R.E., Tindall, J.H., 1983. Effect of postural congruence on client's perception of counselor empathy. Journal of Counseling Psychology 30 (2), 158.
McGarva, A.R., Warner, R.M., 2003. Attraction and social coordination: mutual entrainment of vocal activity rhythms. Journal of Psycholinguistic Research 32 (3), 335–354.
Meltzer, L.E.O., Morris, W.N., 1971. Interruption outcomes and vocal amplitude: explorations in social psychophysics. Journal of Personality and Social Psychology 18 (3), 392–402.
Meltzoff, A.N., Moore, M.K., 1977. Imitation of facial and manual gestures by human neonates. Science 198 (4312), 75–78.
Miles, L.K., Nind, L.K., Macrae, C.N., 2009. The rhythm of rapport: interpersonal synchrony and social perception. Journal of Experimental Social Psychology 45 (3), 585–589.
Mondada, L., 2001. Pour une linguistique interactionnelle. Marges Linguistiques 1, 1–21.
Natale, M., 1975. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. Journal of Personality and Social Psychology 32 (5), 790.
Nenkova, A., Gravano, A., Hirschberg, J., 2008. High frequency word entrainment in spoken dialogue. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. Association for Computational Linguistics, pp. 169–172.
Nishimura, R., Kitaoka, N., Nakagawa, S., 2008. Analysis of relationship between impression of human-to-human conversations and prosodic change and its modeling. In: Ninth Annual Conference of the International Speech Communication Association – INTERSPEECH, pp. 534–537.
Ohala, J.J., 1983. Cross-language use of pitch: an ethological view. Phonetica 40 (1), 1–18.
Oviatt, S., 1996. User-centered modeling for spoken language and multimodal interfaces. IEEE Multimedia 3 (4), 26–35.
Pardo, J.S., 2006. On phonetic convergence during conversational interaction. The Journal of the Acoustical Society of America 119, 2382.
Parrill, 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.
Pentland, A., 2008. Honest Signals – How They Shape Our World. MIT Press, Cambridge.
Pickering, M.J., Ferreira, V.S., 2008. Structural priming: a critical review. Psychological Bulletin 134 (3), 427.
Pickering, M.J., Garrod, S., 2004. Toward a mechanistic psychology of dialogue. The Behavioral and Brain Sciences 27 (2), 169–190 (Discussion 190–226).
Pickering, M.J., Garrod, S., 2006. Alignment as the basis for successful communication. Research on Language and Computation 4 (2–3), 203–228.
Putman, W.B., Street, R.L., 1984. The conception and perception of noncontent speech performance: implications for speech-accommodation theory. International Journal of the Sociology of Language 1984 (46), 97–114.
Ramseyer, F., Tschacher, W., 2010. Nonverbal synchrony or random coincidence? How to tell the difference. Development of Multimodal Interfaces: Active Listening and Synchrony, 182–196.
Richardson, M.J., Marsh, K.L., Isenhower, R.W., Goodman, J.R.L., Schmidt, R.C., 2007. Rocking together: dynamics of intentional and unintentional interpersonal coordination. Human Movement Science 26 (6), 867–891.
Rumsey, F., McCormick, T., 2002. Sound and Recording: An Introduction. Focal Press.
Scherer, K.R., Wallbott, H.G., 1994. Evidence for universality and cultural variation of differential emotion response patterning. Journal of Personality and Social Psychology 66 (2), 310.
Shepard, C.A., Giles, H., Le Poire, B.A., 2001. Communication accommodation theory, The New Handbook of Language and Social Psychology, vol. 1.2, John Wiley & Sons Incorporated. pp. 33–56.
Shockley, K., Baker, A.A., Richardson, M.J., Fowler, C.A., 2007. Articulatory constraints on interpersonal postural coordination. Journal of Experimental Psychology: Human Perception and Performance 33 (1), 201.
Shockley, K., Richardson, D.C., Dale, R., 2009. Conversation and coordinative structures. Topics in Cognitive Science 1 (2), 305–319.
Smith, C., 2007. Prosodic accommodation by French speakers to a non-native interlocutor. In: Proceedings of the XVIth International Congress of Phonetic Sciences, pp. 313–348.
Stanford, G.W., Webster, S., 1996. A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. Journal of Personality and Social Psychology 70 (6), 1231–1240.
Street Jr., Richard L., Street, Nancy James, Van Kleek, Anne, 1983. Speech convergence among talkative and reticent three year-olds. Language Sciences 5 (1), 79–96.

Suzuki, N., Katagiri, Y., 2007. Prosodic alignment in human–computer interaction. Connection Science 19 (2), 131–141.

Tickle-Degnen, L., Rosenthal, R., 2007. The nature of rapport and its nonverbal correlates. Psychological inquiry 1 (4), 285–293.

Van Summers, W., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., Stokes, M.A., 1988. Effects of noise on speech production: acoustic and perceptual analyses. The Journal of the Acoustical Society of America 84, 917.

Vaughan, B., 2011. Prosodic synchrony in co-operative task-based dialogues: a measure of agreement and disagreement. Proceedings of Interspeech 2011. ISCA, pp. 1865–1868.

Vinciarelli, A., 2009. Capturing order in social interactions [social sciences]. IEEE Signal Processing Magazine 26 (5), 133–152.

Ward, Diane, LItman, Arthur, 2007. Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In: ISCA Tutorial and Research Workshop, p. 4.

Ward, N., Nakagawa, S., 2002. Automatic user-adaptive speaking rate selection for information delivery. In: 7th International Conference on Spoken Language Processing (ICSLP2002 – INTERSPEECH 2002).

Webb, J.T., 1972. Interview synchrony: an investigation of two speech rate measures in an automated standardized interview. Studies in Dyadic Communication, 115–133.

Welkowitz, J., Kuc, M., 1973. Interrelationships among warmth, genuineness, empathy, and temporal speech patterns in interpersonal interaction. Journal of Consulting and Clinical Psychology 41 (3), 472–473.

Woodall, G.W., Burgoon, J.K., 1983. Talking fast and changing attitudes: a critique and clarification. Journal of Nonverbal Behavior 8 (2), 126–142.

Zebrowitz, L.A., Brownlow, S., Olson, K., 1992. Baby talk to the babyfaced. Journal of Nonverbal Behavior 16 (3), 143–158.

Zeine, L., Brandt, J.F., 1988. The lombard effect on alaryngeal speech. Journal of Communication Disorders 21 (5), 373–383.

Zhou, J., Su, X., Ylianttila, M., Riekki, J., 2012. Exploring pervasive service computing opportunities for pursuing successful ageing. The Gerontologist, 73–82.

Zuengler, J., 1991. Accommodation in native-non-native interactions: going beyond the 'what' to the 'why' in second language research. In: Contexts of Accommodation: Developments in Applied Sociolinguistics. Cambridge UP, Cambridge, pp. 223–244.