2008-01-01

Towards Measuring Continuous Acoustic Feature Convergence in Unconstrained Spoken Dialogues

Spyros Kousidis
Technological University Dublin, spyros.kousidis@tudublin.ie

David Dorran
Technological University Dublin, david.dorran@tudublin.ie

Yi Wang
Technological University Dublin, yi.wang@tudublin.ie

See next page for additional authors

Funder: the European Union as part of the SALERO project (www.salero.info) through the IST programme under FP6. Part of the research was also funded by Technological University Dublin

Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues

Spyros Kousidis[1], David Dorran[2], Yi Wang[1], Brian Vaughan[1], Charlie Cullen[1], Dermot Campbell[1], Ciaran McDonnell[1] and Eugene Coyle[2]

[1]Digital Media Centre, Dublin Institute of Technology
[2]Audio Research Group, Dublin Institute of Technology

Abstract

Acoustic/prosodic feature (a/p) convergence has been known to occur both in dialogues between humans, as well as in human-computer interactions. Understanding the form and function of convergence is desirable for developing next generation conversational agents, as this will help increase speech recognition performance and naturalness of synthesized speech. Currently, the underlying mechanisms by which continuous and bi-directional convergence occurs are not well understood. In this study, a direct comparison between time-aligned frames shows significant similarity in acoustic feature variation between the two speakers. The method described (TAMA) constitutes a first step towards a quantitative analysis of a/p convergence.

Index terms: Acoustic-prosodic convergence, dialogue speech

1. Introduction

A recent trend in speech technology research is that of studying the phenomenon of *convergence* in spoken dialogues. The term (as used here) refers to acoustic/prosodic (a/p) feature convergence, rather than lexical and/or semantic or cognitive/emotional homonyms [1]. In plain terms, a/p convergence refers to speakers' adaptation of their voice characteristics (such as speech rate, amplitude and pitch), according to those of their dialogue partners. The trend towards studying this phenomenon has been justified in its usefulness both in investigating theories on the collaborative nature of dialogue and behavioural aspects of human communication, as well as in developing highly adaptive and robust spoken dialogue systems and natural sounding speech synthesis components. In the present study, convergence in unconstrained dialogues between two speakers is investigated by use of TAMA (time aligned moving average), a method based on the assumption of *continuous* and *bi-directional* convergence.

Studies from cognitive science have focused on the phenomenon of convergence in dialogues between humans for more than 50 years[1-3]. As highlighted in [2], these studies have evaluated but not quantified the relationship between a/p convergence and the factors they have attributed it to, such as to serve communication efficiency [1, 2], or to express positive evaluation towards the partner [4]. This situation has changed in the past decade, with more studies attempting to quantify relationships between a/p convergence and other semantic/cognitive functions: priming in tutorial sessions [5, 6]; grounding (establishing common ground in discourse) [7]; or expressing and serving communication purposes, such as signalling turn-taking [8]. In addition, a number of studies have investigated convergence in dialogues between humans and conversational agents [9], following a research path from the days of text-based interfaces, when users were found to converge lexically and syntactically to the textual responses of the system [10]. More recently, focus on a/p convergence aims to exploit it in improving performance of dialogue-based interfaces. Automatic speech recognition (ASR) is sensitive to large variations in the acoustic feature vector, so if users adapt their speech to that of the system, then convergence can be utilized to keep the user's speech variation within desirable limits. In addition, a/p features of interactive voice-response systems (IVRs) can be adapted to match those of a random user more closely, as this has been found to sound more pleasant [9, 11].

The majority of these studies use task-based speech corpora, typically simulating the intended application of the system under development, but there are exceptions to this that either use unconstrained dialogue or spontaneous speech [6] corpora. In terms of the specific a/p features studied, the most dominant are speech rate, pause duration (often correlated with turn-taking), and speech amplitude. Finally, diversity among studies exists in terms of time-span and the units involved. Depending on the purpose of the study these can be single vowels, syllables, words, utterances or time frames [2].

This paper presents a methodology which is used to investigate continuous and bi-directional a/p convergence in unconstrained dialogues involving two speakers, by use of a direct comparison of a/p feature averages within time-aligned frames of various sizes. This methodology provides a robust means of measuring convergence of acoustic features (mean pitch, pitch range, intensity, speech rate). For this reason, significant focus has been placed on the corpus acquisition and annotation procedure. Finally, careful consideration has been given so as to avoid assumptions that might bias the analysis, such as arbitrary landmark points (e.g. topic changes) or disregarding speakers' *inherent* speaking styles (in terms of a/p features). For this reason, speaker's a/p features are normalized over their own global averages. In addition, a preliminary analysis of the data acquired using this methodology is presented.

2. Speech corpus acquisition

The dialogue speech analyzed in this study consists of three dialogues with a total duration of 83,7 minutes (average 27,7 minutes per dialogue). The participants were an adult male speaker (speaker A) and three partners: Two adult males (B,

C) and an adult female (D). The participants were aware they were being recorded and were encouraged to engage in 'casual' dialogue. The dialogues took the form of an informal conversation, which included instances of jokes, and other spontaneous dialog acts [12]. Therefore, the recorded speech can be classified as unconstrained dialogue. The reason for this choice is that, although task-based speech corpora are attractive in terms of the limited vocabulary and dialog acts involved, any results cannot be easily generalized; unconstrained dialogue is a more general case where it would be desirable to evaluate the assumption of bi-directional and continuous convergence.

The recording setup (Figure 1) consists of two soundproof isolation booths, equipped with Neumann U87 microphones, Beyer DT150 headphones, computer monitors and Sony CS3N network cameras. These are connected to audio and video consoles which can be operated through dedicated APPLE MAC PRO workstations. The audio is recorded at a very high quality (96KHz/24-bit) in order to ensure lossless sampling of the speech signal as, nowadays, CD quality (44KHz/16-bit) is considered as a minimum standard in audio literature [13]. Optimal recording quality is deemed essential in order to ensure the re-usability of the corpus for future analysis that may benefit from it.

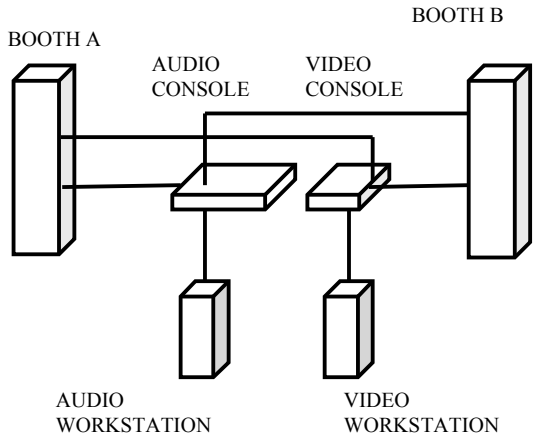

Figure 1 - *Schematic of dialogue recording setup*

The isolation booths offer the advantage of recording each speaker's contribution in a separate track, thus avoiding undesirable artefacts introduced by directional microphone recordings and audio source separation techniques. In addition, the visual feedback can be activated at any time in order to test whether its presence has any effect on a/p convergence, as some of the acoustic cues might be substituted by visual ones. Audio recording is facilitated by the ProTools audio recording application, which is used in professional recording studios. The data is saved in lossless WAV format for further analysis.

3. Analysis method

The recorded audio files are annotated and analysed using the speech analysis program Praat [14]. Each audio file contains the contribution of one speaker. Speech is automatically separated from silence (pauses), based on an intensity threshold. Further corrections are required in order to exclude instances of laughter and other noise (breaths, knocks etc). Each speech interval is then automatically analyzed, and a total of 24 acoustic parameters are measured (mean pitch, mean intensity, minimum and maximum pitch, times of minimum and maximum pitch/intensity, jitter, etc). In addition, vowel detection is performed [15] and the same feature extraction is applied to each vowel. The audio files are separately transcribed and the entire annotation is entered into a multi-purpose online database [16]. The acoustic parameters and vowel enumeration are then used in order to compute the features studied here (see Table 1). The pitch range is expressed as the maximum minus the minimum and the speech rate as vowels per minute.

Feature	Units
Mean Pitch	Hz
Pitch Range	Hz
Mean Intensity	dB SPL
Speech Rate	Vowels/minute

Table 1 - *Acoustic features measured on marked intervals*

In order to make meaningful comparisons between the two speakers, two issues arise. First, for some measures it makes sense to compare them as absolute values (such as speech rate); other measures must be normalized over that speaker's overall mean, as in the case of pitch (comparisons between male and female speakers clearly illustrate this); and for some measures, both normalized and non-normalized comparisons may be meaningful (such as intensity). Since individual speakers have their own inherent speech styles, all of the a/p features are normalized over that speaker's overall mean (see equation 1). A speaker that inherently speaks faster or louder may decrease/increase their tempo or loudness according to similar movements by their partner but not to the extent that they converge in *absolute values*; rather, they may simply probably speak faster/slower than they *usually* do. Therefore, overall means for each feature were calculated using equation 1 (where μ is the overall mean of a feature, f_i is the value of the feature for the interval i, d_i is the duration of interval i, and N is the total number of intervals) in order to compare *normalized* values (i.e. absolute value divided by speaker's overall average for the entire dialogue), in addition to absolute values.

$$\mu = \sum_{i=1}^{N} f_i d_i \Big/ \sum_{i=1}^{N} d_i \qquad (1)$$

In equation (1) above, the feature value (such as mean pitch) is multiplied by the duration of the interval and the overall sum of products is divided by the total speech duration. Thus, speech intervals have a contribution to the overall mean that is proportional to their duration, so that –for example- very short intervals such as back-channelling single word utterances ("uh-um" and "yeah… yeah") need not be excluded from the analysis, due to their inherently lower pitch.

The second issue with comparisons concerns alignment and, in particular, identifying which parts of the dialogues should be compared. An utterance-by-utterance comparison (answer-response) comparison poses two problems: first, to view a dialogue as a series of one-to-one acts and counteracts is rather simplistic and does not adequately represent real dialogue situations [7]; second, some utterance types have inherent local variations in acoustic properties (such as questions having higher pitch than declarative statements). These problems are overcome by using a time-based alignment process: The dialogue is segmented into arbitrary equal-sized frames. The feature values are averaged over the length of the frame, using equation (1). Speech intervals that cross over frame boundaries are clipped-off at those

boundaries (see Figure 2) without re-measuring the a/p features. This results in a reduction in speech interval duration, which in turn equals the proportional contribution of each interval to that frame's average.

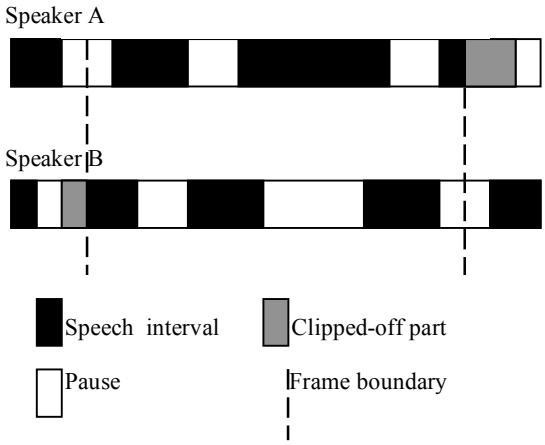

Figure 2 - *Schematic of dialogue frame*

The average of a feature for a frame corresponds to a number of utterances and other non-lexical dialog acts (such as back-channelling sounds, "uh-ums" etc). The frames can be overlapping, a technique that resembles a Moving Average filter, in that it causes a smoothing of the resulting contour. Thus, this process is referred to here as TAMA (time aligned moving average). The resulting feature averages of frames for the two speakers are simultaneously plotted in a scatter plot as points, which are connected by smooth lines (see Figure 3).

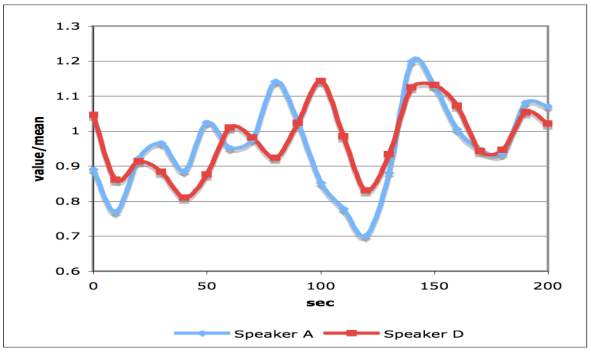

Figure 3 – *Average normalized speech rate for speakers A, D over 20 second frames with 50% overlap*

Four different frame sizes are used (10,20,30 and 60 seconds), in order to investigate the time span of convergence. Frames of small duration are likely to contain a single utterance (or part thereof), thus resembling an utterance-based analysis. Thus, it was predicted that the types of utterances (declarative vs. interrogative, back-channelling) within a given frame would determine the degree of convergence. Incrementally longer frames gradually show a more 'global' trend, as they are more likely to contain several utterance types and therefore tend to be more representative of the true feature average. Using too large a frame length introduces the risk of the frame average not being significantly different from the overall average.

4. Results and discussion

The overall feature averages calculated with equation 1 are shown in Table 2. The "total length" column refers to the total duration of speech intervals analyzed for that speaker (not including pauses).

Speaker	Total length (s)	Mean Pitch (Hz)	Pitch Range (Hz)	Mean Intensity (dB SPL)	Speech Rate (vowels /min)
A	868	107	90	47	213
B	933	126	127	48	236
A	762	128	74	71	221
C	868	125	73	73	243
A	736	110	79	60	217
D	633	167	132	61	189

Table 2 - *Overall Feature Averages*

The most significant convergence was found for Intensity. This is independent of the frame length, which suggests that convergence of amplitude occurs promptly and that speakers readily adjust their "volumes" to a mutually "agreed" level. However, dissimilarities in movements such as those indicated by the dashed rectangles in Figure 4 regularly occur. The first irregularity (left) is attributed to the fact that, within those frames, speaker C keeps the turn for most of the time, while A's average is based mostly on a sudden question, characterized by higher intensity, and a few non-lexical very short utterances ("wow"). The second irregularity is a result of A speaking and laughing at the same time. Notably, this was not removed from the analysis at the manual annotation stage, as it is a perfectly intelligible utterance, although the speaker is laughing. Longer frame settings have a "smoothing" effect and show even higher convergence between the two speakers' intensity averages.

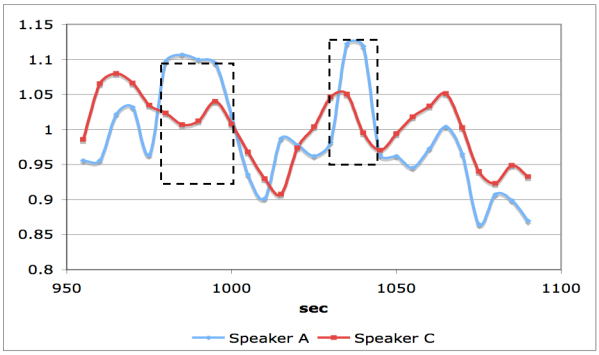

Figure 4 - *Average normalized intensity for speakers A, C over 10 second frames with 50% overlap*

Speech rate was the feature that showed the second most significant convergence (see Figure 3). Unfortunately, inaccuracy in automatic vowel detection and pauses *within* utterances introduce significant error in the calculation of speech rate. However, longer frames (which are likely to contain a more balanced time-share and intra-sentence pause

duration between the two speakers) better illustrate convergence, as the errors 'cancel out' each other. In the future, this problem will be dealt with by introducing intra-sentence pause annotation in the corpus, as well as manual correction of the automatic vowel detection, so that speech rate convergence can be measured more accurately.

Average pitch (Figure 5) was also found to converge, but less significantly than speech rate. This is because pitch serves several functions and different utterance types have largely different pitch configurations: Back-channelling word-long expressions generally have low pitch, while expressions of enthusiasm ("Wow") have very high pitch; Prosodic functions such as interrogative vs. declarative tone, and focal stress (word or sentence) also have a significant effect on pitch. Part-of-speech tagging is required in order to overcome this problem. This will allow TAMA analysis of specific types of utterances, and it is believed that a convergence trend will be shown more accurately for average pitch.

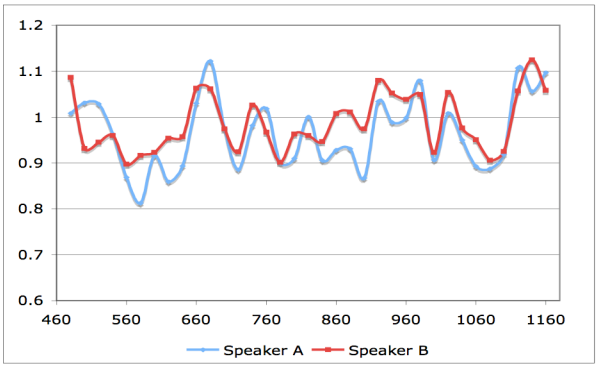

Figure 5 - *Average normalized pitch for speakers A, B over 30 second frames with 33% overlap*

Pitch range was not found to converge significantly in any of the dialogues analyzed here. However, pitch range measurement is prone to errors because of octave jumps that occur when using the pitch detection algorithm of Praat [14]. In the future, the measurement of pitch range will be made more accurate by detecting and extracting values from the pitch contours. In addition, pitch range also heavily depends on utterance type. Therefore, part-of-speech tagging may provide insights into pitch range convergence.

Overall, convergence was not found to change over time, as the dialogue progresses. Rather, speakers appear to converge promptly and not deviate, as long as they keep exchanging turns equally. Where one of the speakers keeps the turn for a long time, the other speaker's contribution is typically reduced to back-channelling and a/p convergence is masked by the quiet speaker's silence. It is noted that no significant difference was found in convergence of both speech rate and intensity as absolute values, rather than normalized, although the normalization method (division over the overall mean) is *not* equivalent to *standardisation* (i.e. z-scores).

5. Conclusions and future work

An approach to measuring a/p convergence in unconstrained dialogues is proposed. This preliminary study shows convergence of intensity and speech rate. Less significant convergence was found for mean pitch and no significant convergence was found for pitch range. Further work is required to investigate convergence of those features. In addition to measuring convergence trends more accurately, this will allow modelling of a/p convergence of specific utterance types (such as back-channelling), which is useful for dialogue-based applications. Further, it is intended that more a/p features be studied in terms of convergence, such as measures of voice quality, stress patterns, syllable duration and variations of intensity (intensity range). Focus on a/p convergence will be useful for development of conversational agents that exhibit more natural a/p "behaviour", both in terms of "understanding" (ASR) as well as "communicating" (speech synthesis).

6. Acknowledgements

A substantial part of this research was funded by the European Union as part of the SALERO project (www.salero.info) through the IST programme under FP6. Part of the research was also funded by Dublin Institute of Technology (www.dit.ie).

7. References.

[1] M. J. Pickering and Simon Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences, 27: 169-190 Cambridge University Press,* vol. 27, pp. 169-190, April 2004.

[2] S. Oviatt, C. Darves, and R. Coulston, "Toward adaptive conversational interfaces: Modeling speech convergence with animated personas," *ACM Trans. Comput.-Hum. Interact.,* vol. 11, pp. 300-328, 2004.

[3] N. Suzuki and Y. Katagiri, "Prosodic alignment in human-computer interaction," *Connection Science,* vol. 19, pp. 131 - 141, 2007.

[4] J. Welkowitz and M. Kuc, "Interrelationships Among Warmth, Genuineness, Empathy, And Temporal Speech Patterns In Interpersonal Interaction," *Journal of Consulting And Clinical Psychology,* vol. 41, p. 472, 1973.

[5] A. Ward and D. Litman, "Dialog Convergence and Learning," in *13th International Conference on Artificial Intelligence in Education*, Los Angeles, CA, 2007.

[6] D. Reitter, Johanna D. Moore, and F. Keller, "Priming of syntactic rules in task-oriented dialogue and spontaneous conversation.," in *28th Annual Conference of the Cognitive Science Society (CogSci)*, Vancouver, Canada, 2006, pp. 685-690.

[7] I. Mushin, L. Stirling, J. Fletcher, and R. Wales, "Discourse Structure, Grounding, and Prosody in Task-Oriented Dialogue," *Discourse Processes,* vol. 35, pp. 1 - 31, 2003.

[8] L. t. Bosch, N. Oostdijk, and L. Boves, "On temporal aspects of turn taking in conversational dialogues," *Speech Communication,* vol. 47, pp. 80-86, 2005.

[9] L. Bell, J. Gustafson, and M. Heldner, "Prosodic adaptation in human-computer interaction," in *ICPhS 2003*, Barcelona, 2003, pp. 2453-2456.

[10] E. Zoltan-Ford, "How to get people to say and type what computers can understand," *Int. J. Man-Mach. Stud.,* vol. 34, pp. 527-547, 1991.

[11] N. Ward and S. Nakagawa, "Automatic User-Adaptive Speaking Rate Selection," *International Journal of Speech Technology,* pp. 259-268, October, 2004 2004.

[12] H. F. Wright, "Modelling prosodic and dialogue information for automatic speech recognition." vol. PhD: Unviresity of Edingburgh, 1999.

[13] R. A. Katz, *Mastering Audio: The Art and the Science*: Focal Press, 2002.

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 4.4.18 ed, 2006.

[15] W. H. Press, S. A. Teukolsky, and W. T. Vetterling, *Nimerical Recipes in C: The Art of Scientific Computing*: Cambridge University Press, 1992.

[16] C. Cullen, B. Vaughan, and S. Kousidis, " Emotional Speech Corpus Construction, Annotation and Distribution," in *The 6th edition of the Language Resources and Evaluation Conference* Marrakech (Morocco), 2008.