2009-01-01

Monitoring Convergence of Temporal Features in Spontaneous Dialogue Speech

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

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

Funder: 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.

Monitoring Convergence of Temporal Features in Spontaneous Dialogue Speech

Spyros Kousidis[1] *and David Dorran*[2]

[1] Digital Media Centre, Dublin Institute of Technology, Ireland
[2] Audio Research Group, Dublin Institute of Technology
spyros.kousidis@dit.ie

Abstract

This paper presents ongoing research on convergence of speech features in human dialogues, in view of simulating this behaviour in spoken dialogue systems. The TAMA method (time-aligned moving average), previously used on monitoring convergence of acoustic prosodic (a/p) features, is applied to temporal properties of speech (between-turn pauses and overlaps). The results are compared to those of an older study on the same features.
Index Terms: convergence, spoken dialogue systems

1. Introduction

Spoken dialogue systems (SDS) present an attractive interface for many applications, as speech is the simplest and most efficient type of communication. For certain applications, it is beneficial for SDS to be viewed through a *human metaphor* [1], as if the users were talking to a person, rather than a machine. Increasing the naturalness of the interaction is one way to enhance this metaphor, hence the need for more "human-like" behaviour on the part of the SDS. As suggested in [1], convergence of speech features between two interlocutors, a property of human dialogue that is well-known in behavioral and communication sciences [2, 3] may well increase the perceived naturalness of an SDS, if the latter can exhibit the same behaviour. Towards this end, convergence in human dialogues must be well understood.
Following previous work on monitoring convergence of acoustic/prosodic (a/p) features in human dialogues [4], the work presented here focuses on temporal aspects of dialogue, and in particular the duration of between-turn pauses and overlaps. The analysis follows a method similar to that adopted in [5], with the addition of the TAMA (time-aligned moving average) method [4].

2. Speech convergence

The phenomenon of convergence (sometimes termed alignment, entrainment or accommodation) refers to an observed behaviour in human dialogues, mainly a tendency of the interlocutors to "match" or "converge" in certain properties of their speech [2]. These properties are numerous and include lexical accommodation (using common terms without explicitly agreeing to do so), accent, dialect or pronunciation (typical in dialogs among members of the same ethnic/cultural group), acoustic/prosodic features (F0, intensity, pitch range) and temporal features (pauses, speech rate, overlaps).
A variety of theories exist that attribute convergence to different functions and can be categorized into four main groups [6]: (a) *biological models*, that specify adaptation to the interlocutor as an autonomous, spontaneous response, (b) *arousal and affect models,* which view convergence (or absence thereof) as an affective response, therefore redefining the behaviour as habitual, (c) *social norm models,* where situational context and social background are explanatory factors for displayed habitual/intended behaviour and, (d) *communication and cognition models,* where communication accommodation is seen as a conscious strategy in some cases.
Whether convergence is seen as a habitual or mechanistic [3] response, the fact remains that it is a property of human dialogue that persists even when one of the interlocutors is replaced by a conversational interface [7, 8]. Humans tend to adapt their speech to that of a conversational interface. This has been exploited to some extent in SDS and IVRs (interactive voice response systems), as convergence of speech rhythm on the part of the SDS was found to be positively evaluated in [9], while in [10], ASR performance was improved by keeping the users' speech rate within specific limits, as they unknowingly adapted to the speech rate of the system.
However, the phenomenon of convergence is much more complex than the engineered solutions imply, and further analysis of human dialogue corpora is required in order to properly implement this behaviour in SDS design. One particular issue is multimodality, i.e. the fact that interlocutors may converge along one or several "dimensions"[2] (speech properties) *simultaneously.* Therefore, investigation of several dimensions is required (both individually and in parallel) in order to understand the process more adequately. Hence, this paper presents the application of the TAMA method, previously used in the analysis of convergence of a/p features, in order to investigate convergence of temporal features.

3. Data acquisition and feature extraction

The speech corpus for this study consists of five dialogues (8 different speakers) recorded during a task-based application scenario. Subjects are situated in soundproof isolation booths and communicate without visual contact to each other. The scenario requires the subjects to rank 15 items (Figure 1 - Screenshot from dialogue recording experiment (Himalayas scenario)1) in order of importance within a limited amount of time, so as to survive a hypothetical hazard, such as a shipwreck, being stranded in space, or being lost in the Himalayas. The contributions from each speaker are recorded in separate audio channels. The speech segments are automatically detected (using an intensity threshold) by a Praat script [11] and manually corrected for detection errors. The resulting *chronograph* (Figure 2) of the dialogue contains the required turn-switching information.

3.1. Turn switching

A uniform definition for "turns" in dialogues is lacking. In this study, the definition used in [5] has been adopted: each speaker's speech segments are processed separately. Immediately before a speech segment, there is either a pause

or an "overlap" segment. If a pause is found, then the speech segment before the pause is examined. If that segment belongs to the *other* speaker, then the pause is "between turns". A pause is attributed to the speaker that starts speaking *after* the pause (see Figure 3).

Figure 1 - Screenshot from dialogue recording experiment (Himalayas scenario)

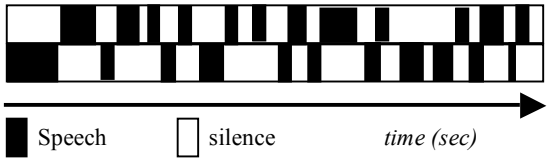


Figure 2 - Chronograph of turn taking between two speakers

3.2. Continuous monitoring of pause length and overlap rate

In order to monitor the evolution over time of the pause length and overlap rate, the TAMA method [4] is used. Synchronous overlapping frames of fixed length are applied to each speaker's chronograph. For each frame, the average between-turn pause length and amount of overlaps are computed. The *overlap rate* is defined as the amount of turns that begin with an overlap (rather than a pause), over the total number of turns in that frame. In [5], overlaps were not attributed to speakers, due to ambiguity in resolving which speaker they belong to in some cases. Here, the same rule that was applied to the pause length was used: overlaps belong to the speaker that *keeps* the turn after the overlap, as it was found that ambiguous cases are rare. In the example shown in Figure 3, the average pause length (APL) for speaker A is p(2) (the length of pause 2), and for speaker B the APL is (p(1)+p(3)+p(6))/3. The overlap rate (OR) for speaker A is 0.5 (one overlap in two turns), and the OR for speaker B is 0.25 (one overlap in four turns).

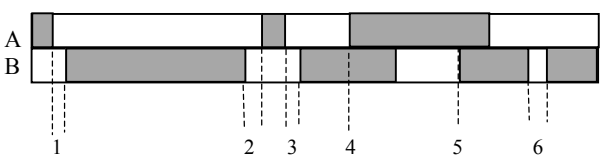


Figure 3 - Schematic of TAMA frame for pause length and overlap rate computation. According to adopted definition, (1),(3) and (6) are pauses before turns of speaker B, (2) is a pause before a turn of speaker A, (4) is overlap switch to A, and (5) is overlap switch to B.

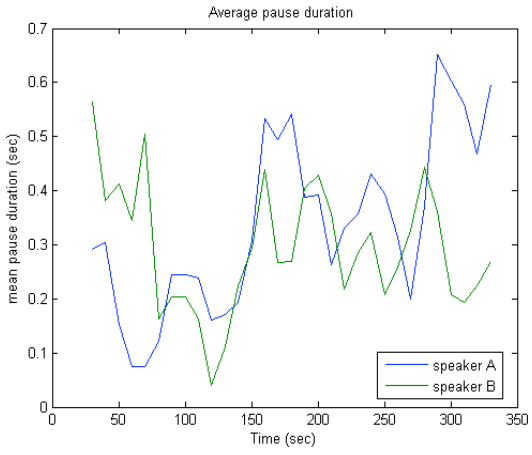
4. Results and Discussion

Overall averages for pause length and overlap rate are shown in Table 1. A significant correlation is found for average pause length (R=0.96, p=0.008). In [5] it was reported that this is the result of pause length accommodation (convergence) between the two speakers. The TAMA plots (see Figure 4) reveal a slightly different picture. These plots were obtained by applying the TAMA method with a frame length of 30 seconds and an overlap of 20s (simulating a 3-point simple moving average). One can discern a similar trend between the two speakers in both plots, pointing to the convergence hypothesis. However, such trend similarity is more often than not indiscernible in the TAMA plots (e.g. Figure 5), and even more often statistically insignificant. This is either the result of absence of convergence in some parts of the dialogue, or an indication that the analysis is too simplistic; it is very likely that different "modes of dialogue" or utterance types (such as back-channeling) have different turn-switching configurations. Therefore, the average of any frame will be a function of convergence (if present) and the specific characteristics of the interaction at that time. Hence, An SDS design strategy of converging to the user's overall average seems inadequate, as that would produce a response from the system that displays much less variation than that of a user and may well seem unnatural for some dialogue situations and/or utterance types.

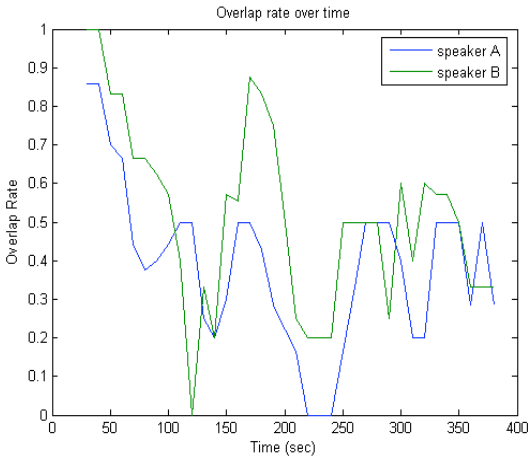
D	Speaker A			Speaker B		
	TT	OR	APL (sec)	TT	OR	APL (sec)
1	95	0.44	0.77	82	0.44	1.03
2	120	0.63	0.31	119	0.59	0.25
3	121	0.62	0.28	109	0.53	0.33
4	87	0.41	0.49	69	0.53	0.46
5	71	0.49	0.39	69	0.33	0.32

Table 1- Results from analysis of five dialogues. (TT = total turns, OR = overlap rate, APL = average pause length)

For overlap rate, the overall averages do not show high correlation, but the trend similarity in plots such as Figure 4(b) points to similar conclusions as for the pause length. Due to the small sample available, these results need to be validated with more experiments.

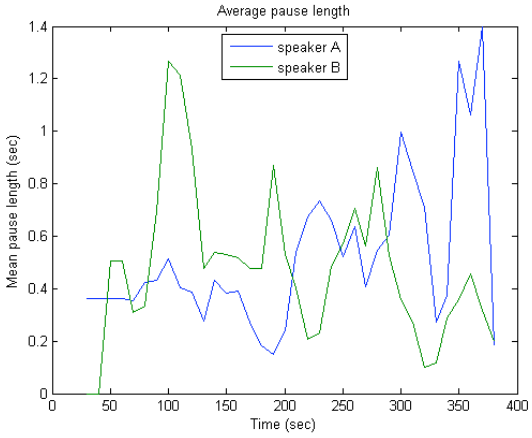


(a)

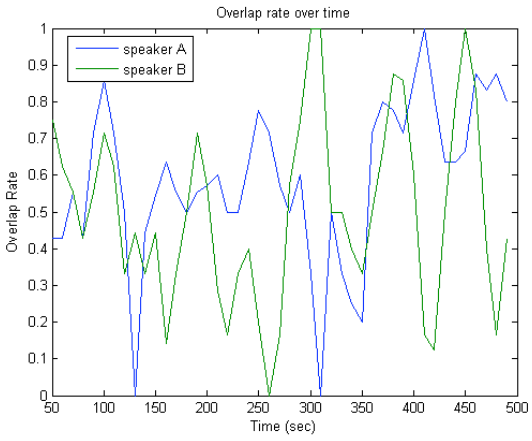


(b)

Figure 4 - TAMA plots produced by calculating averages for overlapping frames (frame length=30s, overlap=20s) (a) Average pause length (b) Overlap rate. Speakers appear to follow a similar trend in both plots.



(a)



(b)

Figure 5 - TAMA plots (frame length 30s, overlap 20s) (a) Average pause length (b) Overlap rate. Speakers do not appear to follow a similar trend in either plot

5. Conclusions and future work

The preliminary results show that convergence in pause length and overlap rate is worth investigating. The convergence hypothesis holds true for the overall average between-turn pause length, but not for individual TAMA frames. Further wok is required to investigate a more complex model for convergence of temporal features. The goal of this work is the implementation of such a model in a prototype SDS, in order to evaluate whether (and to what extent) the interaction is perceived as more "human-like".

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] J. Edlund, J. Gustafson, M. Heldnera, and A. Hjalmarsson, "Towards human-like spoken dialogue systems," *Speech communication,* vol. 50, pp. 630-645, 2008.

[2] H. Giles, A. Mulac, J. J. Bradac, and P. Johnson, "Speech Accomodation Theory: The First Decade and Beyond," in *Communication Yearbook 10*, M. L. McLaughlin, Ed. Newbury Park: SAGE, 1987, pp. 13-48.

[3] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences,* vol. 27, pp. 169-190, April 2004.

[4] S. Kousidis, D. Dorran, Y. Wang, B. Vaughan, C. Cullen, D. Campbell, c. McDonnell, and E. Coyle, "Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues " in *Interspeech 2008* Brisbane, Australia, 2008.

[5] L. t. Bosch, N. Oostdijk, and L. Boves, "On temporal aspects of turn taking in conversational dialogues," *Speech Communication,* vol. 50, pp. 80-86, 2005.

[6] J. K. Burgoon, L. A. Stern, and L. Dillman, *Interpersonal Adaptation: Dyadic Interaction Patterns*: Cambridge university Press, 1995.

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

[8] 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.

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

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

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