Measuring prosodic alignment in cooperative task-based conversations

Khiet P. Truong, Dirk Heylen

Human Media Interaction, University of Twente, Enschede, The Netherlands

{k.p.truong,d.k.j.heylen}@utwente.nl

Abstract

In this paper, we investigate prosodic alignment in task-based conversations. We use the HCRC Map Task Corpus and investigate how familiarity affects prosodic alignment and how task success is related to prosodic alignment. A variety of existing alignment measures is used and applied to our data. In particular, a windowed cross-correlation procedure, that has been used previously in visual behavior research, is applied to prosodic features. In addition, we address the issue of how to separate genuine observed alignment from alignment that is a result from random coincidental behavior. Using these measures, we find some indications of prosodic convergence and synchrony in the map task conversations. Alignment tendencies are strongest for intensity, and familiarity seems to play a role in convergence. Finally, weak evidence was found for a correlation between prosodic alignment measures and task success.

Index Terms: prosodic alignment, convergence, synchrony, familiarity

1. Introduction

According to the Speech Accommodation Theory [1], people accommodate their speech behaviors to each other in conversation. This is presumably (unconsciously) done to create rapport and a positive harmonious atmosphere. Some studies have also shown that alignment is positively correlated with task success: for example, it was found that entrainment in high-frequency words [2], and lexical and syntactic repetition [3] are predictive of task success. These studies showed that automatic measures of lexical alignment correlate with task success. Based on [4], it is assumed that alignment on one level boosts alignment on other levels. Hence, we are interested to see whether there is also *prosodic* alignment present in task-based conversation. Measuring prosodic alignment requires a somewhat different approach than measuring lexical alignment.

Recent works on (automatically) measuring prosodic alignment include the so-called TAMA method, proposed by [5]. It is based on a 'time-aligned moving average': by calculating moving averages of the acoustic features under investigation, a visual inspection of alignment is facilitated. However, it was not explained how alignment could be quantitatively measured. TAMA seems to be a popular measuring method given that several studies have used this method to quantify speech alignment. It was used in [6] in combination with a coupled oscillators model and the authors concluded that speech similarity changes during social interaction. A similar conclusion was drawn by [7] who used TAMA in a windowed correlation procedure. Other studies have used more linguistically meaningful units instead of windows with a certain size. Prosodic alignment was locally quantified in [8] by addressing turn changes, and by computing alignment measures between each consecutive turn. These measures were then successfully used in a classification task of positive versus negative attitude in married couples' interactions. In another study [9], correlations between acoustic features extracted from adjacent turns were computed and it was concluded that these features showed 'proximity' and synchrony at the turn level. In [10], it was suggested to use measurement methods that can capture dynamic temporal aspects of alignment. Alignment of gaps and pauses was measured by first applying some pre-processing to these features to transform the discontinuous nature of the durations of gaps and pauses into continuous feature streams. This process allowed a comparison between two speakers' speech features at any possible timestamp.

A somewhat scattered view on the evidence of prosodic alignment processes in conversations and on how to actually measure prosodic alignment emerges from the studies described. Evidence for prosodic convergence and synchrony were relatively small, and were usually shown for a small number of conversations. But the evidence was also not conclusive; for some studies there was strong prosodic alignment found for a certain feature but not in another study. All studies acknowledge that a dynamic approach to alignment should be undertaken – most of the studies use a moving window approach. It would be interesting to combine this moving window approach with a certain latency to see whether alignment is led or followed by certain persons, as suggested by e.g., [10]. Another issue that has not frequently been discussed in works on prosodic alignment is the matter of how to separate 'real' speaker-specific speech alignment processes from random coincidental speech behaviors. This issue was touched upon in [9] by pairing a target speaker with another randomly chosen person other than the original interlocutor, and by looking at whether the acoustic differences between this fabricated pair would be smaller or larger than the original pair of speakers. Although this is an important issue in alignment research, it seems to have been much more of a subject of study in bodily and gestural behavior-based alignment research (see e.g., [11, 12]) than it has been in speech-based alignment research.

We will attempt to address some of the aspects mentioned in the works reviewed. Particularly, we will focus on a dynamic approach to measure prosodic alignment, we will address the 'coincidental alignment' issue, we will investigate whether familiarity plays a role in alignment, and we will look at whether the alignment measures considered here correlate with task success. In the remainder of this paper, the word 'alignment' will be used to cover a broad range of phenomena that have something to do with 'adapting one's speaking behavior to another one's speaking behavior'. We will also use more specific terms such as convergence and synchrony which we adopt from [10].

The paper is structured as follows. Section 2 briefly describes the HCRC Map Task corpus used for this analysis. In Section 3, we give a description of the alignment measures considered in this study. The results are presented in Section 4. We

conclude with a discussion and a few words on future research in Section 5.

2. Data

For our analysis, we used the HCRC Map Task Corpus [13] that consists of Scottish English spoken task-based dyadic conversations held under various conditions. These conditions involve whether there is eye contact or not, and whether the participants in the conversations are familiar (FAM) with each other or not (UNFAM). We were interested in the familiarity dimension and decided to use 31 FAM[1] and 32 UNFAM conversations from the no-eye-contact condition (out of the 128 available conversations). Since we were interested in vocal alignment, we used only the no-eye-contact condition and expected that the effect of vocal alignment would be more apparent when effects of visual behaviors are ruled out (evidence for this was already found in [14] where it was illustrated that face-to-face interactions show more and longer simultaneous speech than in non-face-to-face interactions, suggesting less synchronicity). Each participant was assigned a certain role, that of a giver or follower. The task was to enable the follower to reproduce the giver's route on the follower's map. The maps contain certain landmarks and differ between each giver and follower. Task success was measured in terms of how far the route that the follower has drawn deviates from the route shown on the giver's map[2].

3. Analysis

We adopt the concepts of convergence and synchrony as defined in [10] where the process of convergence is described as 'two parameters becoming more similar over time'. Synchrony is described as 'parameters/events happening at the same time or working at the same speed'.

3.1. Feature processing

The first step was to create so-called talkspurts from the continuous speech stream to have some workable units. The silence/speech classification used was provided by the manual transcription available in the corpus. Using these classifications, silences of less than 200 ms were bridged by speech, and speech events shorter than 100 ms were bridged by silence in order to create talkspurts. Log F_0 and intensity were measured continuously with a time step of 0.01 s using Praat [15]. For an analysis of convergence and synchrony, a meaningful pairing between the speakers' feature values is necessary which is complicated by the fact that our speech features are discontinuous and misaligned between the two speakers. Log F_0 and intensity for speech analysis only make sense when there is speech involved and this speech usually does not occur at the same time for both speakers. Therefore, all features were transformed to a continuous feature stream. With respect to F_0 and intensity: averages over each talkspurt were taken, followed by an overlapping moving window that averages over 6 data points (i.e., 6 talkspurts or 6 gaps or 6 pauses; this was mainly done to smooth the contour), followed by a linear interpolation between the averages obtained. All speech features were transformed to z-scores. For convergence analysis of intensity, we also report the non-transformed intensity value as we wanted to see whether people align on intensity in an absolute or relative way. For synchrony analysis, the z-transformation of intensity did not change the relative behavior of the non-transformed intensity, so only results from the raw intensity measurements are reported.

[1] Conversation q3nc3 was discarded due to microphone problems

[2] These path deviation scores are included in the 2.1 release of the corpus' annotations.

3.2. Convergence

For convergence, we adopt similar procedures and measures as proposed in [10]. These measures have in common that they intend to capture the decreasing difference (in time) of a certain feature between two speakers. The first measure concerns a simple Pearson correlation between the differences of the two speakers' feature values and the time – the more negative the correlation, the stronger the convergence. For the second measure, all conversations were divided into equally-sized first and second halves. The difference between the feature's mean measured over these two halves gives an indication of whether the participants have become 'closer' to each other towards the end of the conversation. For convergence, this difference between these two halves should be positive (the 2nd halve is subtracted from the 1st halve), and it should be significantly different between the two halves.

3.3. Synchrony

For measuring prosodic synchrony, we adopt a windowed cross-correlation (wcc) procedure, originally proposed by [16] and which has been applied to visual movement synchrony [16, 12]. This method is suitable for capturing the dynamics and locality of speech synchrony as it takes into account possible lags in processes of synchrony: it allows an analysis of leading and following speech behaviors in time. The method is based on a windowed correlation procedure (i.e., Pearson correlation is calculated for each overlapping moving window). Extending this method to a cross-correlation procedure means that during each window, additional correlations are computed over a pair of signals that are shifted with respect to each other by certain lags in time (forward or backward). There are several parameters that need to be chosen by the researcher. The window size (of the window that is moved along the signal) should be chosen large enough such that correlations can be reliably computed, but small enough to capture the dynamicity. We chose a window size of 20 s that moved across the signal with a time step of 10 s. The maximum lag and the increment of this lag determines how much and how often one of the paired feature vectors is shifted forward or backward. We chose a maximum lag of -20 and 20 s and an increment size of 5 s. The results of this windowed cross-correlation procedure can be given in a results matrix where each cell represents the correlation between two signals, of which one of them can have a certain lag, measured over a certain window size at a specific time. This matrix can be visualized as shown in Fig. 1. For a more detailed description and the exact computation of the windowed cross-correlation procedure, readers are referred to [16].

3.4. Coincidence or not?

Several approaches have been proposed in previous research to rule out the possibility that the amount of convergence or synchrony found is caused by random coincidence. The general idea behind these approaches is to generate 'pseudointeractions' – if the alignment found in real interactions is genuine, it should be stronger in real interactions than in pseudointeractions. Pseudointeractions can be generated in different ways. In [11], the following was proposed: in order to generate a pseudointerac-

tion for a real interaction AB between speakers A and B, take A and B from additional real interactions AC and DB to generate a 'fake' interaction 'AB'. Unfortunately, for most of the conversational speech corpora, this is not a feasible method (due to the fact that most corpora have speakers that only talk to an interlocutor once). Therefore, we propose a method that draws from [11, 12] to generate pseudointeractions that yield more realistic comparisons and conservative testing. Each interaction is divided into 5 equally-sized segments (proportionally to the duration of the interaction). Recall that we applied linear interpolation to the moving averaged measurements taken over 6 talkspurts. For a real interaction AB, we select a random speaker X as 'fake' B. We use genuine B's timestamps of the averaged measurements prior to interpolation and generate random measurements at those timestamps to produce a 'fake' B to be paired with A. With respect to synchrony, these random measurements are constrained by the rule that they have to be drawn from the same time segment as where the real measurement occurred. In other words, B's feature value at timestamp t in time segment 2 must be replaced with one of X's feature values shuffled within X's time segment 2. This is done to keep the timing structure somewhat intact (avoiding A's data point timed near the beginning to be paired with 'fake' B's value timed in the end of the conversation for example). Subsequently, linear interpolation is performed. This procedure is repeated 10 times for each speaker A and B such that each real interaction can be compared to 20 corresponding pseudointeractions. With respect to convergence, this timing constraint was discarded because the ordering structure plays a role in convergence.

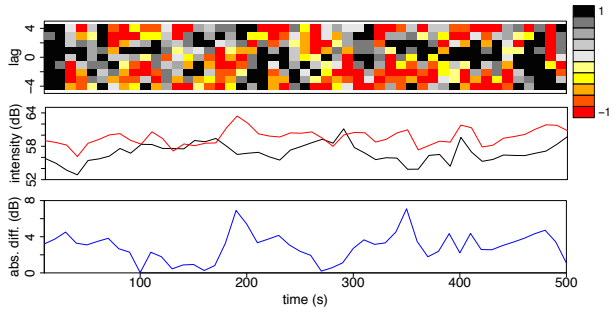

Figure 1: *A visual representation of the wcc method applied to one of the conversations of the HCRC corpus. The top pane shows the correlations obtained with the wcc procedure: the y-axis shows the lag and the x-axis the time. Black colors show positive correlations while red colors show negative correlations. The middle pane shows the smoothed intensity contours. The bottom pane shows the absolute difference between the two intensity contours.*

4. Results

4.1. Convergence

The results for convergence are shown in Table 1. In general, the amount of convergence found is relatively low. Absolute and relative intensity show signs of convergence. There are no significant differences between the UNFAM and FAM conditions but there are tendencies indicating that people seem to converge more in the UNFAM situation than in the FAM situation (given the significant and larger mean differences in intensity and the stronger negative correlations between time and the absolute differences for the UNFAM condition). One could speculate that people who are unfamiliar with each other show more pronounced convergence behavior because they have to get to know each other while people who are familiar with each other already have gone through that process.

We compared the measures obtained with the real interaction to the measures obtained with the pseudointeractions. It seemed that the results obtained with the real interactions are not significantly different from the pseudointeractions which makes it difficult to draw conclusive conclusions from these results although tendencies are visible.

Table 1: *Convergence results. * means that the averaged differences between the 1st and 2nd halves are statistically significant at $p < 0.05$ (one-sided paired t-test). Standard deviations are given in brackets. Numbers in bold mean significantly higher values than pseudointeractions ($p < 0.05$)*

Feature	UNFAM	FAM
	Mean diff.: 1st minus 2nd half	
Intensity	0.75 (2.25)*	0.53 (2.42)
Intensity_z	**0.18 (0.36)***	0.064 (0.22)
F_0_z	-0.04 (0.08)	-0.06 (0.09)
	Correlation between time and abs. diff.	
Intensity	-0.10 (0.36)	-0.05 (0.41)
Intensity_z	**-0.18 (0.39)**	-0.06 (0.41)
F_0_z	0.11 (0.13)	0.13 (0.16)

4.2. Synchrony

A visual representation of the wcc procedure for one of the conversations is shown in Fig. 1. This figure allows for a visual inspection of the dynamics of alignment, and hence we believe that such figures can be very useful for a more detailed analysis. Table 2 shows the results for synchrony. In general, the strength of synchrony found is relatively low. We can observe that synchrony is more pronounced for intensity than F_0_z.

Furthermore, there does not seem to be a significant difference between the UNFAM and FAM condition (except in one case). The results obtained were compared with pseudointeractions. Paired t-tests showed that most of the pseudointeractions yielded synchrony levels that were significantly lower ($p < 0.01$) than the synchrony levels of the real interactions, indicating that people do show speaker-specific behavior to some extent.

Table 2: *Synchrony results with several measures. * means that* UNFAM *differs signifantly from* FAM *at $p < 0.01$. Standard deviation are given in brackets. Numbers in bold mean significantly higher values than pseudointeractions ($p < 0.05$).*

feature	UNFAM		FAM
	static Pearson		
Intensity	**0.23 (0.31)**	*	**0.13 (0.28)**
F_0_z	0.13 (0.24)		**0.15 (0.22)**
	windowed		
Intensity	**0.12 (0.19)**		**0.10 (0.15)**
F_0_z	**0.07 (0.15)**		0.04 (0.20)
	wcc max		
Intensity	0.84 (0.06)		**0.85 (0.06)**
F_0_z	**0.86 (0.05)**		0.84 (0.07)

4.3. Correlation with task success

In order to see whether task success is influenced by the amount of convergence and/or synchrony, we looked for correlations between our measures and the path deviation scores that are an indication of task success: the lower the path deviation score, the larger the success. The correlations are shown in Table 3. With respect to convergence, intensity_z shows a relatively weak correlation with path deviation score (note the direction of correlation that points towards a positive relationship between task success and a certain measure of alignment, indicated by arrows in Table 3). With respect to synchrony, a relatively weak positive relationship (the more synchrony, the lower the path deviation score) was found for intensity as well. To see whether a combination of convergence and synchrony measures would yield stronger relations between alignment and task success, we carried out a multiple regression with the convergence and synchrony measures based on intensity_z as the 4 predictor variables and the path deviation score as the dependent variable – an R-squared of 0.13 was found.

Table 3: *Correlations between convergence and synchrony measures, and task success (measured over both* UNFAM *and* FAM*). P-values that approach statistical significance are shown in brackets. Arrows indicate whether a positive or negative correlation indicates a positive relationship between task success and a certain measure of alignment.*

	Intensity	Intensity_z	F_0_z
Convergence – mean diff ↘	-0.09	-0.22 (p=0.09)	0.04
Convergence – corr. between time and abs. diff. ↗	0.08	0.32 (p=0.01)	-0.06
Synchrony – static Pearson ↘	-0.19		0.07
Synchrony – windowed ↘	-0.06		0.07
Synchrony – wcc max ↘	-0.24 (p=0.06)		0.02

5. Discussion and conclusions

We have presented several methods and measures to quantify prosodic alignment in terms of convergence and synchrony. The results obtained showed tendencies towards convergence and synchrony. Alignment effects were more pronounced for intensity than for F_0. Familiarity seems to have an effect on alignment but this observation needs further investigation. Task success seems to be weakly related to the alignment of (relative) intensity. In addition, we proposed a way to rule out the possibility that the obtained results were due to random coincidence. We believe that these kinds of tests are necessary to show that the observed alignment is really a result of speaker-specific adaptation.

The measurement of alignment remains a complicated matter, partly due to its dynamic nature and the social factors that influence the amount of alignment. We have tried to capture these dynamics through a windowed cross-correlation procedure which introduces lags along a moving window. However, how to represent and quantify these dynamics remains a challenge. The visualization of the wcc procedure as shown in Fig. 1 presents a start.

Future research should concentrate on the dynamics of alignment and take time lags into account. Lags were taken into account in this study but we did not further analyze leading or following behaviors which could give us insights into the social dynamics between the speakers.

6. Acknowledgements

We would like to thank three anonymous reviewers for their helpful comments. This research has been supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet).

7. References

[1] H. Giles, D. M. Taylor, and R. Bourhis, "Towards a theory of interpersonal accommodation through language: some canadian data," *Language in Society*, pp. 177–192, 2010.

[2] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of ACL/HLT*, 2008, pp. 169–172.

[3] D. Reitter and J. D. Moore, "Predicting success in dialogue," in *Proc. 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 808–815.

[4] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, pp. 169–226, 2004.

[5] 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 *Proceedings of Interspeech*, 2008, pp. 1692–1695.

[6] De Looze, C. and Rauzy, S., "Measuring speakers' similarity in speech by means of prosodic cues: methods and potential," in *Proceedings of Interspeech*, 2011, pp. 1393–1396.

[7] B. Vaughan, "Prosodic synchrony in co-operative task-based dialogues: A measure of agreement and disagreement," in *Proceedings of Interspeech*, 2011, pp. 1865–1867.

[8] C.-C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. G. Georgiou, and S. Naryanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Proceedings of Interspeech*, 2010, pp. 793–796.

[9] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proceedings of Interspeech*, 2011, pp. 3081–3084.

[10] J. Edlund, M. Heldner, and J. Hirschberg, "Pause and gap length in face-to-face interaction," in *Proceedings of Interspeech*, 2009, pp. 2779–2782.

[11] F. J. Bernieri and R. Rosenthal, "Interpersonal coordination: Behavior matching and interactional synchrony," in *Fundamentals of nonverbal behavior*, R. S. Feldman and B. Rime, Eds. New York: Camgbridge University Press, 1991, pp. 401–432.

[12] F. Ramseyer and W. Tschacher, "Nonverbal synchrony or random coincidence? How to tell the difference," in *COST 2102 International Training School 2009*, A. Esposito, Ed. Heidelberg: Springer Verlag, 2010, pp. 182–196.

[13] A. H. Anderson, M. Bader, E. Gurman Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weintert, "The HCRC Map Task Corpus," *Language and Speech*, vol. 34, pp. 351–366, 1991.

[14] D. R. Rutter and G. M. Stephenson, "The role of visual communication in synchronizing conversation," *European Journal of Social Psychology*, vol. 7, pp. 29–37, 1977.

[15] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[16] S. M. Boker, M. Xu, J. L. Rotondo, and K. King, "Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series," *Psychological Methods*, vol. 7, pp. 338–355, 2002.