A STUDY OF PROSODIC ALIGNMENT IN INTERLINGUAL MAP-TASK DIALOGUES

Hayakawa Akira, Loredana Cerrato, Nick Campbell, Saturnino Luz

School of Computer Science and Statistics, Trinity College Dublin, Ireland
campbeak@scss.tcd.ie, cerratol@tcd.ie, nick@tcd.ie, luzs@cs.tcd.ie

ABSTRACT

This paper reports results from a study of how speakers adjust their speaking style in relation to errors from Automatic Speech Recognition (ASR), while performing an Interlingual map task. The dialogues we analysed were collected using a prototype speech-to-speech translation system which adds 3 elements to the communication which we think of as "filters": Automatic Speech Recognition (ASR), Machine Translation (MT) and Text To Speech (TTS). Our belief is that these filters affect the speakers' performance in terms of cognitive load, resulting in adaptation of their communicative behaviour. The study shows that the participants do adjust their speaking style and speaking rate as a way of adapting to the errors made by the system. Specifically, the results show that (a) system errors influence speaking rate, and (b) the perceived level of cooperation by the interlocutors increases as system error increases.

Keywords: Interlingual communication, map task, speaker alignment, speech rate, adaptation.

1. INTRODUCTION

While the commercial deployment of large-scale automatic speech-to-speech translation systems is now becoming a reality, empirical data showing the constraints imposed by this interlingual context are still scarce. To fill this gap and to better understand strategies for adaptation to language technology components we carried out a data collection in a specific environment designed for the purpose of investigating how Automatic Speech Recognition (ASR), Machine Translation (MT) and Text To Speech (TTS) synthesis affect users in terms of cognitive load, adaptation of communicative behaviour to the technology, and repair strategies.

Using a prototype system able to record synchronised interaction data streams, such as high quality video and audio, time-stamped ASR, MT and TTS events as well as biosignals (heart rate, skin conductance, blood volume pressure and EEG) we collected a corpus of 15 dialogues between interlocutors who speak two different languages (English and Portuguese) [4]. The dialogues were elicited using the Edinburgh Map task technique [1].

Although a number of studies of linguistic phenomena and adaptation strategies in map task dialogues have been carried out [8, 5] to the best of our knowledge, this is the first investigation of communicative behaviour in map task dialogues in a computer mediated interlingual setting.

Previous research on cultural difference and adaptation of communicative styles in computer mediated intercultural communication has focused on identifying and categorising the types of problems that arise in intercultural dialogues. This is in order to apply machine learning techniques to coded dialogues with the aim of automatically recognising when problems arise (or are likely to arise) in intercultural conversations [13].

In our study the focus is on how interlocutors adjust their speaking behaviour in terms of speaking style and speaking rate to adapt to the errors made by the system which mediates their interaction. Spoken dialog can be seen as a joint action in which interlocutors coordinate their verbal and non-verbal behaviour and adapt their linguistic choices to each other. This often results in the phenomenon referred to as convergence or alignment which consists in a tendency by the interlocutors to adopt and re-use each other's language patterns in the course of authentic interaction [2]. According to the interactive alignment model, originally proposed by Pickering & Garrod [7] this linguistic coordination in dialogue occurs at the level of the lexicon, grammar, and pronunciation and represents one way in which interlocutors achieve understanding in dialogue.

Convergence is a property of human dialogue that seems to persist even when one of the interlocutors is replaced by a conversational interface [6, 12]. Given that interlocutors adapt their linguistic choices to each other and given that humans tend to adapt their speech even to that of a conversational interface, our question is if some kind of adaption still occurs when two interlocutors are placed in a setting where their interaction is mediated by a translation system. In particular in this experiment we observed the be-

haviour of two interlocutors who could not see each other since they were sitting in two different rooms, their interaction was not transmitted via video, and they could not hear each other since their content was mediated by an interlingual system with a synthetic voice as final output. The aim of the study was to investigate whether the ASR performance would have an effect on adaptation. Specifically, we were interested to determine the extent to which the distribution of errors in the ASR affect the speaking style in terms of hyperarticulation and speech rate and the complexity of the structure of the utterances.

2. MATERIALS AND METHOD

By using a prototype interlingual communication system [4], we collected a corpus of task-based dialogues between speakers of two different European languages (English and Portuguese). The corpus includes: high quality audio of the participants' utterances, video, eye tracking, physiological signals (EEG, BVP, SC), and ASR, MT, TTS events which are synchronised and finely time-stamped. These dialogues were elicited with the map task technique [1, 3] which is still the ideal way of eliciting natural conversation in a controlled situation given the simplicity of the task and the complexity of phenomena it can elicit. Our map task corpus follows the original design of the HCRC Map Task Corpus, but is based on speech-to-speech machine translated interaction, where the two dialogue participants, the instruction giver and the instruction follower, speak different languages. The instruction giver has a map with a route drawn on it and s/he has to instruct the follower to draw the route on his/her unmarked copy of the map. Neither participant can see the other's map. Each map contains a number of reference points (e.g., "white mountain", "baboons" "waterfall"). Some features are common to both maps, and some differences between the reference points are incorporated in the maps in order to make the dialogues more complex.

In our interlingual setting, turn taking is quite systematic since there is a push-to-talk button in the system. The sessions were recorded in two settings, one in which the participant could see each other (Video-On) and one in which they could not (Video-Off). Neither participants can hear the other's voice since the output of the ASR and MT is provided by a synthetic voice. For this study we analysed a subset of seven dialogues between English and Portuguese speakers, in which the interlocutors could not see each other (Video-Off).

The dialogues were orthographically transcribed with the addition of some labels for interruptions, filled and empty pauses and noises. Transcription of the dialogues was carried out manually by two students (one native speaker of English and one native speaker of Portuguese) who listened to the audio-channel using Wafesurfer [10]. The transcribers were also asked to judge whether the interlocutors were interacting with each other or more interacting with the system and whether they where behaving in a cooperative way. The answers were given on a seven-point Likert scale going from 1 strongly disagree to 7 strongly agree. Since in the seven analysed dialogues the participants cannot hear each other's voice, and since it has been shown that humans tend to adapt their speech even to that of a conversational interface [6, 9], we suppose that they align their speaking rate to that of the synthetic voice (180 wpm).

To calculate deviation in the speaking rate we made a comparison between the duration of the utterances of the actual participant and the duration of the output of the given utterance by the same TTS (which acted as a reference) system used during the recordings. Speech recognition performance was assessed by aligning the reference transcripts and the ASR-generated hypothesis and computing word error rate (WER), precision, recall and f-score per sentence. WER is defined as the ratio of the Levenshtein distance between the aligned utterances (i.e. the number of additions, substitutions and deletions needed to convert one of the utterances into the other) to the number of words in the reference transcript. Precision (P) is the ratio of matching words in the alignment to the number of words in the hypothesis. Recall (R) is the ratio of matches to words in the reference, and f-score (F_0) is the harmonic mean of P and R.

Finally, we assessed possible correlations between the distribution of the errors (WER) and the behaviour of the interlocutors in terms of graded interactivity and cooperation in the dialogues. Our hypotheses were (1) that higher WER would make the interaction between the interlocutors more difficult and thus make it seem as though the interlocutors were interacting with the system, rather than with each other, and (2) that when the interaction gets "more difficult" (because of repeated error) a series of repair strategies is used by the interlocutors to make the ASR work better, thus either shifting the interactive behaviour from interlocutor to the the system or generating cooperative behaviour to compensate for the errors.

In order to test these hypotheses, we macro-averaged the WER per dialogue, discretised the er-

ror distribution by categorising the scores as low, medium, high and very high WER, and compared them to the transcribers' scores for interactivity and cooperation (7-point Likert scale) using the Kendall rank correlation coefficient statistic.

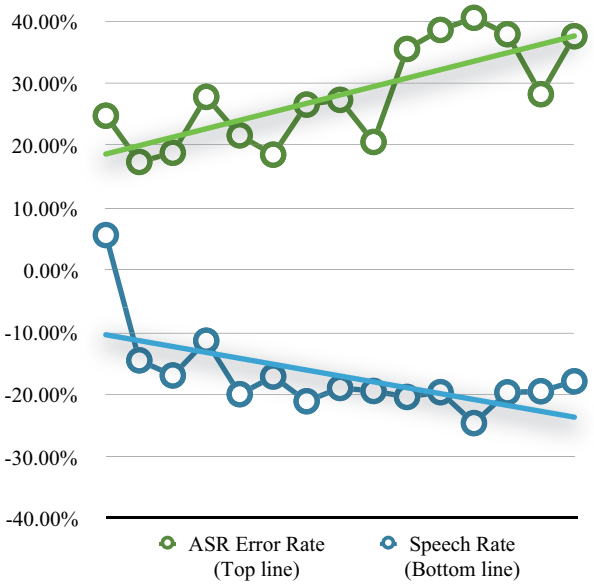
3. RESULTS

We calculated the duration of the utterances of each interlocutor and compared it to the TTS utterance duration – so that the TTS speaking rate acts as a reference to observe deviation in the interlocutors speaking rate related to a measure of the accuracy of the ASR performance. The results of the first 3 quarters of all seven analysed interlocutors are plotted in Figure 1, where in the blue (bottom) curve we can observe the deviation in speaking rate:

- 0% means that the interlocutor's utterance duration is the same as the TTS utterance duration.
- Positive percentage points indicate that the interlocutor's utterance is faster than the TTS.
- Negative percentage points indicate that the interlocutor's utterance is slower than the TTS.

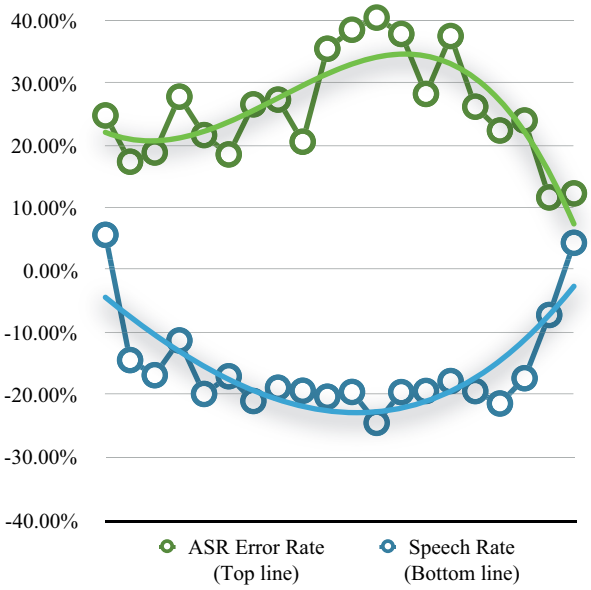
Thus, the farther away from 0%, the faster (+ figures) or slower (- figure) the interlocutor's utterance is. The top (green) curve represents a measure of the ASR performance accuracy (inverted f-score) – 0% if the ASR is completely correct, and 100% if the ASR was completely wrong.

Figure 1: Deviation of speaking rate related to ASR accuracy for all seven speakers (first three quarters of dialogue – Q1 to Q3).



Although the recovery strategy was quite different across users we can observe a clear trend: soon after errors occur the speakers tend to slow down their speaking rate, but this effect dies out after a while and they either return to their normal speaking rate or they speak faster. It's interesting to notice that on a whole, the speaking rate becomes slower as the task continues, but the ASR accuracy does not improve. Besides speaking in a slower way the speakers tend to hyperarticulate and often repeat key words or produce utterances with simple syntactic structure. However when we look at the whole dialogue, as plotted in Figure 2, we see in the forth quarter the speaking rate increases when the task nears or reaches the end, with an improvement to the ASR accuracy – apart from speaking faster, the speaker still continues to repeat key words and produce utterances with simple syntactic structure.

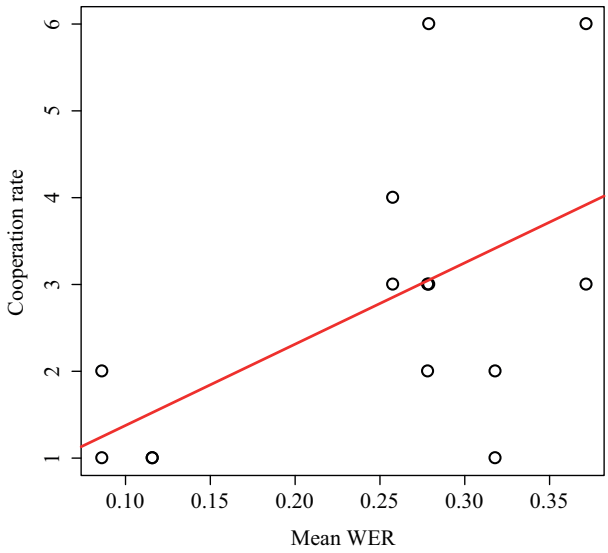
Figure 2: Deviation of speaking rate related to ASR accuracy for all seven speakers (whole dialogue – Q1 to Q4).



As for the correlation between the distribution of the errors (WER) and the behaviour of the interlocutors in terms of graded interactivity and cooperation in the dialogues, using the Kendall correlation test we get no apparent correlation between WER and interactivity ratings ($\tau = -0.15$, $p < 0.69$, non-significant). The data suggests, as one would expect, a negative correlation (i.e. the higher the WER, the less the transcribers thought the speaker was interacting with the other, rather than the system). However the correlation is quite weak, and not statistically significant. One might speculate that this is due to the fact that the question posed to the transcribers was not specific enough, in addition to the small size of the data set.

However, even in this small sample we were able to detect a correlation between WER and the cooperation ratings ($\tau = 0.66$, $p < 0.05$). Figure 3 shows the correlation plot of WER by cooperation rating, with a linear model fit. This correlation is positive and could be interpreted as indicating that the higher the WER the harder the speaker tries to behave cooperatively. This means speaker uses different repair strategies such as: speaking more clearly and slower (hyperarticulating), repeating key words, producing shorter utterances with simple syntactical structure and so on.

Figure 3: Correlation between WER and rated cooperation in the dialogues.



4. DISCUSSION

The result shows that speaking rate and clear speech (hyperarticulation) are used as repair strategies when ASR errors occur. This behaviour is not a generalised and stable mode of speaking in the dialogues we analysed, since it seems to be a targeted and flexible adaptation strategy. This is in line with the results obtained in a previous study carried out by Stent et al. [11] who showed evidence for a relation between misrecognition and hyperarticulation in computer-directed speech. We observe a similar behaviour in our dialogues, however the novelty in our results consists in the fact that they show for the first time evidence of adapting behaviour and repair strategies even in a multimodal interlingual map task setting where the speech is not exactly computer directed, but human directed mediated by a system.

Moreover in our study the misrecognition are authentic and not simulated. Our corpus represents a valid source of useful data that can provide novel contribution towards a deeper understanding of device mediated interlingual contexts.

5. CONCLUSIONS AND FUTURE WORK

The results of an initial investigation aiming at describing how speakers adjust their speaking style in relation to errors from Automatic Speech Recognition (ASR), while performing an Interlingual map task are here reported and discussed. Given the complexity of the phenomenon of convergence, we are interested in studying several other aspects at different levels: phonetic, lexical, syntactic and also observe non-verbal behaviour and cognitive load by analysing the physiological signals.

We are therefore performing annotation of several communicative phenomena on the data in our corpus. Our goal for the future is to carry out further analysis of different phenomena related to convergence in this interlingual map task setting. The aim of which is to gain a deeper understanding of how communication works in this setting and hopefully implement a newer improved version of the system which takes into account specific aspects of interlingual communication.

Beside the traditional analysis of speech, gestures, and facial expressions we plan to investigate possible correlations between the participants? brain activity (EEG), blood volume pulse and skin conductance during points of difficulty within the interactions (for instance, when amused, frustrated or surprised).

6. ACKNOWLEDGEMENTS

This research is supported by the Science Foundation Ireland (grant 07/CE/I1142) as part of the CNGL Centre for Global Intelligent Content (www.cngl.ie) and ADAPT (www.adaptcentre.ie) at Trinity College, Dublin.

We would also like to thank Tim Ryan, Cesar Sabato for their contribution to this corpus and Helen Türk for some helpful comments.

7. REFERENCES

[1] Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., Weinert, R. Oct. 1991. The hcrc map task corpus. *Language and Speech* 34(4), 351–366.

[2] Giles, H., Mulac, A., Bradac, J., Johnson, P. 1987. Speech accommodation theory: The first decade and beyond. In: L.McLaughlin, M., (ed), *The First Decade and Beyond, in Communication Yearbook 10*. Newbury Park: Sage 13–48.

[3] Gorisch, J., Astésano, C., Bard, E. G., Bigi, B., Prévot, L. 2014. Aix map task corpus: The french multimodal corpus of task-oriented dialogue. *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* Reykjavik. ELRA 2648–2652.

[4] Hayakawa, A., Campbell, N., Luz, S. 2014. Interlingual map task corpus collection. *Proc. of 15th INTERSPEECH* Singapore. ISCA 189–191.

[5] Newlands, A., Anderson, A. H., Mullin, J. 2003. Adapting communicative strategies to computer-mediated communication: an analysis of task performance and dialogue structure. *Applied Cognitive Psychology* 17(3), 325–348.

[6] Oviatt, S., Darves, C., Coulston, R. 2004. Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *ACM Transactions on Computer-Human Interaction (TOCHI)* 11(3), 300–328.

[7] Pickering, M. J., Garrod, S. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27(02), 169–190.

[8] Reitter, D., Moore, J. D. 2007. Predicting Success in Dialogue. *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics* Prague. Association for Computational Linguistics 808–815.

[9] Schneider, A., Luz, S. 2011. Speaker alignment in synthesised, machine translated communication. *International Workshop on Spoken Language Translation* San Francisco. ISCA 254–260.

[10] Sjölander, K., Beskow, J. 2000. Wavesurfer - an open source speech tool. *Proc. of the 6th International Conference on Spoken Language Processing* Beijing. ISCA 464–467.

[11] Stent, A. J., Huffman, M. K., Brennan, S. E. 2008. Adapting speaking after evidence of misrecognition: Local and global hyperarticulation. *Speech Communication* 50(3), 163–178.

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

[13] Wang, H.-C., Fussell, S. F., Setlock, L. D. 2009. Cultural difference and adaptation of communication styles in computer-mediated group brainstorming. *Proc. of the SIGCHI Conference on Human Factors in Computing Systems* Boston. ACM 669–678.