TURN-TAKING PREDICTIONS ACROSS LANGUAGES AND GENRES USING AN LSTM RECURRENT NEURAL NETWORK

Nigel G. Ward, Diego Aguirre, Gerardo Cervantes and Olac Fuentes

University of Texas at El Paso
Computer Science Department
500 West University Avenue, El Paso, TX 79968, USA

ABSTRACT

Going beyond turn-taking models built to solve specific tasks, such as predicting if a user will hold his/her turn after a pause, there is growing interest in more general models for turn taking that subsume many such tasks, and very good results have recently been obtained [1]. Here we present an improved recurrent network model that outperforms [1] and does so without requiring lexical annotation. Further, we show that this model can be trained for different languages with no modifications, providing good results in turn-taking prediction for English, Spanish, Japanese, Mandarin and French. We also show that our model performs well across genres, including task-oriented dialog and general conversation.

Index Terms— Turn-taking models, Spoken dialog systems, LSTM

1. INTRODUCTION

Participants in a conversation generally take turns speaking. For spoken dialog systems, good turn-taking is similarly necessary for efficient and natural-feeling interactions [2]. This involves, for example, avoiding or at least minimizing interruptions, unnecessary pauses, and overlaps. To accomplish this, every spoken dialog system employs some kind of turn-taking model, although in most deployed systems these are very basic.

Research on turn-taking has demonstrated that it is possible to do much better [3]. Indeed, specific turn-taking decisions — such as when to start a turn and when to produce a backchannel — can, with sufficient training data, be made with accuracies and at speeds that match or exceed human performance [4, 5, 6]. However, each of these demonstrations has been costly to build, requiring a major engineering effort, starting with collection and preparation of extensive data. Moreover, each of the resulting systems has, in the end, been demonstrated successfully only for one specific decision, although ultimately we want systems to that perform well in response to a variety of user behaviors.

One specific limitation is that models of turn-taking have traditionally been trained to make decisions at discrete time-points — for example, times after detecting speaker pauses of some fixed duration — to decide whether he/she intends to continue or to yield the turn. Over the past two years, however, several research groups have applied Long Short-Term Memory (LSTM) networks [10] to turn-taking [1, 11, 12, 13, 14, 15, 16]. LSTM networks seem well-suited to turn-taking since they can track the situation continuously as it evolves over time while representing long-distance context effects. Skantze's work [1] in particular showed good performance as a general model and also, without further training, performed well on specific traditional tasks. Remarkably, it achieved better-than-human performance on the classic turn hold/take discrimination task.

In this paper, taking Skantze's work as a starting point, we present a better model, and use this model to explore the predictability of turn taking in five languages and three genres.

2. MODEL

A general model of turn taking should be able not just to make one kind of turn-taking decision, but instead to predict the probability that a given participant will speak over a future time window [1, 17]. In particular, a model should continuously predict future behavior at every time step, not just when certain events occur.

We accordingly operationalize the task, following Skantze, as one of predicting, at each moment, for each speaker, whether he or she will be speaking in the immediate future. The most relevant scope of prediction (prediction horizon) depends on the application and technology. For many dialog systems this is on the order of one second, that being the time it can take from making a decision, to speak or stop speaking, to the time when that decision is realized. However this can be much shorter with more incremental systems [6, 18, 19]. Here we do experiments for five prediction windows: from 0 to 250ms, 0 to 500ms, 0 to 1s, 0 to 2s, and 0 to 3s.

We thank the Toyota Motor Corporation and DARPA's Lorelei program for support. No endorsements should be inferred.

2.1. Features

For prediction we use a number of prosodic features and related measures, 6 for each speaker, as follows: 1) *Absolute Pitch*, in log Hz, computed using Voicebox [20] 2) *Relative Pitch* (z normalized per track). Both pitch features have values of 0 for frames where no pitch was detected. 3) *Voicing*, based on the pitch-detector output, is a binary feature indicating whether the frame was voiced or not. 4) *Energy*, not normalized. 5) *Voice activity* is a binary feature indicating whether the participant was speaking at the moment. This feature was not computed, but obtained from the human generated labels. 6) *Cepstral flux* is the proxy measure for lengthening. The value of this feature is low when a phoneme is lengthened, that is, when the sound changes little from one frame to the next, and high when the speaking rate is high. The code for these features is freely available [21].

These 6 features are very similar to those of Skantze. However our cepstral flux measure is approximately the inverse of Skantze's spectral stability metric. We decided to use cepstral coefficients, rather than frequency band power values since they may better approximate perceptions of the spectral envelope. While Skantze's best results were obtained using in addition human-generated part of speech tags, in many practical uses such information is unavailable or unreliable, and so the model we build does not use such features.

2.2. Architecture

In overview, our model takes as input 12 low-level features extracted from the speech activity of both speakers, 6 per speaker, all computed every 50ms. These features are fed to the network in a stream as they become available. The heart of the network is a LSTM layer. Its output is an n-dimensional vector, where n is the number of 50-millisecond frames in the prediction window. This is then passed through a final layer to produce the network's output. Each of the values in this final n-dimensional vector represents a prediction of whether the speaker will speak (1) or not (0) during that future frame. In these respects our model is the same as that of Skantze. The rest of the section describes the improvements we made.

Figure 1 gives an overview of our architecture. The input is first processed by a dense layer of Parametric Rectified Linear Units (PReLU) [22], with 12 hidden units. In this layer, to improve robustness and reduce overfitting, we added dropout [23] with a *keep* probability of 0.75. This process is repeated a second time by another pair of dense and dropout layers with the same parameters and properties. The output of the second dropout layer is then fed to a LSTM layer with 30 hidden units. The output layer is a dense layer, where the PReLU operation is applied to each of the n prediction values produced by the LSTM layer. Since the values produced by the PReLU operation are not necessarily in the range of [0,1], their outputs are clipped to force them to be in that range. Thus the key differences from Skantze's model include the use of more hidden LSTM units, PReLU at various layers, including at the output layer instead of a sigmoid, and the use of more layers, especially dropout layers. In preliminary small-scale experiments we found all of these to be beneficial.

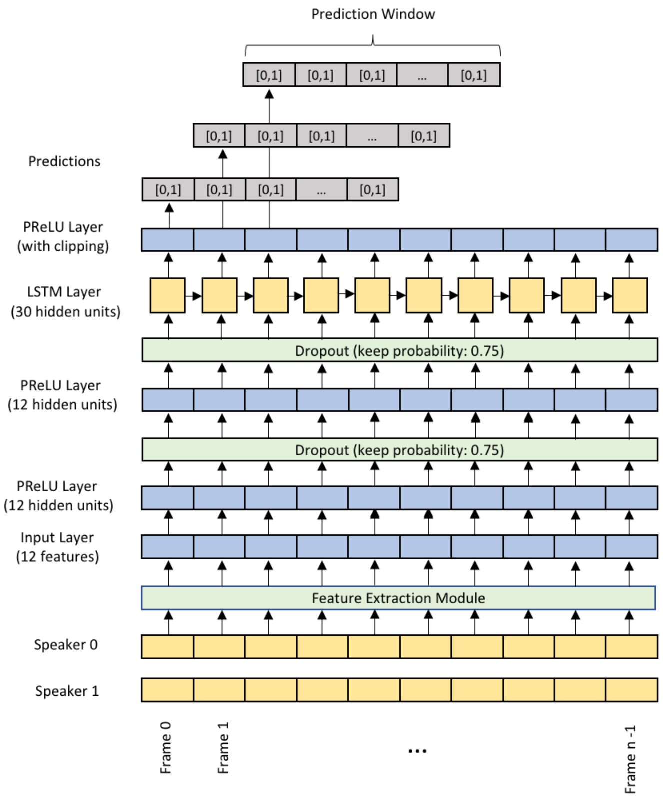

Fig. 1. Proposed Network Architecture

The code for this model is freely available [24].

2.3. Training

We used Tensorflow for model construction and evaluation. The training data was partitioned into mini-batches of size 128, each with a sequence length of 1200 frames (60 seconds). We used a learning rate of 0.001 and to train the networks for 1,200 epochs using the RMSProp optimizer. We chose as our loss function the mean-squared error (MSE) between the ground truth, from the labels, and our predictions before clipping. We used an L2 regularization factor of 0.001. We trained a different network for each time window size (0 to 250ms, 0 to 500ms, and so on). In each case we used the mean absolute error (MAE) metric as the measure of the performance of the network on the test set after every epoch. For each window, the MAE between the prediction and the ground truth is computed as the average of the mean absolute error across all points in our prediction window, that is, from the current time frame (the point of prediction) to the last frame in the window (at the prediction horizon). This resembles Skantze's training procedure, except for the use of larger mini-batches, a smaller learning rate, and backpropagation without truncation through time. Although all these changes increased the computation required, it still takes only 1 to 2 hours to train a network in most cases, depending on the data size, on a desktop with two GPUs.

3. EVALUATION

This section describes a direct comparison of our model with that of Skantze. To do this, as Skantze's code is not publicly available, we first reimplemented his model. Apart from using our own feature extractors, we did this faithfully. We then did a head-to-head comparison using the same dataset Skantze used, namely the original Maptask data [25]. We used an 80:20 training-testing split. Table 1 shows Skantze's reported results and our own results, but for our reimplementation of his model and for our own model. In addition, Figure 2 compares the MAE of our reimplementation and our own model as a function of training time.

Prediction	Skantze's Model		Our
Window	Original	Reimplemented	Model
0 to 250ms	0.15	0.14	**0.11**
0 to 500ms	0.22	0.23	**0.19**
0 to 1s	0.28	0.32	**0.27**
0 to 2s	**0.33**	0.37	0.34
0 to 3s	**0.35**	0.39	0.36

Table 1. Mean Absolute Error (MAE) — English Maptask

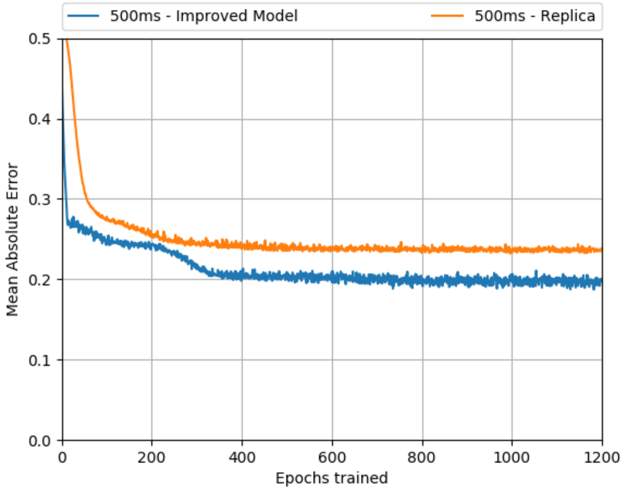

Fig. 2. 500ms Model Training Results — Maptask

We observe several things. 1) The results of our reimplemntation were close to those of Skantze, but generally somewhat inferior, most likely due to the omission of a part-of-speech feature. 2) Our architecture did substantially better than that of Skantze, showing the advantages of our improvements. 3) In the prediction horizons of greatest practical interest, 0–500 ms and 0–1 s, our model performs better than Skantze's, even without the part-of-speech feature. 4) Prediction is much easier for closer time windows.

An important advantage for general models is the ability to perform well on more specific problems, without being specifically trained for them. To demonstrate this, we test the ability of our model to predict whether, after a brief pause, a speaker will hold his/her turn or whether the turn will shift to the other speaker. Following Skantze, we found all the points where there was a pause (250ms/500ms) after someone stopped speaking. At these points, to convert our framewise predictions into a single hard decision, we took the predictions for the next second, and then we averaged these for each side. The speaker with the highest prediction average was then the one we predicted would take the turn. In this procedure we exactly followed Skantze, again to enable a direct comparison to his reported results. These are seen in Table 2. The results are mixed, and in part reflect different operating points, but for the 500ms case our model is doing roughly as well as Skantze's, again without using part of speech tags.

4. THE PREDICTABILITY OF TURN TAKING ACROSS LANGUAGES

While the patterns of turn-taking are known to have some commonalities across languages, a model trained on one language cannot in general be expected to perform well on another [26, 27]. That is, to build a good turn-taking model for any new language we can expect to need to train it on data from that language. In this section we estimate the ability of our model to learn and represent the patterns of turn taking in different languages.

Our first comparison is between the English and Japanese Maptask corpora [28]. In this section our metric is the percent reduction in mean absolute error from a baseline model that always predicts silence; this baseline of course varies from corpus to corpus Table 3 shows the performance of the model on these two data sets. We see that our model also performs well for Japanese, although the benefit is much less.

This relates to a larger question, of how languages differ in the predictability of turn-taking, and in particular how much value is provided by prosodic information. Only one previous study seems to have addressed this question, that by Brusco and colleagues, who found slightly better performance for hold versus shift predictions in American English than in Argentine Spanish, using the same model [27]. Similarly, the difference we found can be taken to suggests that turn-taking in Japanese involves prosody less than it does in English. This is somewhat surprising, given evidence that backchanneling in Japanese is more prosodically-controlled

	after a 250ms pause			after a 500ms pause		
Instances	3,405	7,546		2,079	4,608	
% Hold	59.8%	58.8%		57.6%	57.6%	
Model	Skantze	Replica	Ours	Skantze	Replica	Ours
Shift: Precision	0.726	0.776	**0.784**	0.711	0.780	**0.800**
Shift: Recall	**0.703**	0.528	0.601	**0.738**	0.549	0.660
Shift: F-measure	**0.714**	0.628	0.680	**0.724**	0.644	0.720
Hold: Precision	**0.805**	0.730	0.759	**0.802**	0.727	0.778
Hold: Recall	0.822	**0.893**	0.884	0.780	**0.886**	0.879
Hold: F -measure	0.813	0.803	**0.817**	0.791	0.799	**0.825**

Table 2. Hold/Shift Prediction Results — MapTask

than in English [29], but perhaps this can be understood as reflecting a difference in how languages allocate the available prosodic bandwidth to different functions.

Prediction	English		Japanese	
Window	MAE	% reduction	MAE	% reduction
0 to 250ms	0.11	67%	0.21	38%
0 to 500ms	0.19	42%	0.30	12%
0 to 1s	0.27	18%	0.36	-6%
0 to 2s	0.34	-3%	0.40	-18%
0 to 3s	0.36	-9%	0.41	-21%
baseline	0.33		0.34	

Table 3. Mean Absolute Error — English and Japanese Maptask

Our second comparison is across five languages using corpora of telephone speech from the Callhome/Callfriend collection, for which we used the provided training/testing splits. Although American English Callhome is far larger, the results do not change appreciably when only a comparable amount of data is used. The results are shown in Table 4. Again we see that the model performs well across languages, but again less well for Japanese, suggesting that its turn taking is harder to predict from prosody alone.

5. THE PREDICTABILITY OF TURN TAKING ACROSS GENRES

For Japanese we experimented with one additional data set. This was collected in a companion robot scenario, in which the robot was controlled by a remote hidden operator, Wizard of Oz style (WoZ). This was different from the other data collections in that it included long response delays and many misunderstandings rather than smoothly flowing interactions, and in that the human participants were senior citizens rather than college students. Tables 5 and 6 summarize our results on this data set. We observe that the predictions of shifts were quite poor. We suspect that this is due to two factors: first the unbalanced nature of this data set, and second the fact the generally isolated comment-response pairs in this data set were a poor match for the LSTM models' implicit assumption that the dialog state is flowing continuously.

From Tables 3, 4 and 5, we observe large differences in performance for the three Japanese corpora. One possible factor may be the level of familiarity and resulting differences in the formality of turn-taking: in the telephone corpus the interactants are friends or family members, in the Maptask they were fellow students, and in the WoZ data they were strangers.

Incidentally, for all languages and all window sizes, our architecture generally outperformed Skantze's. There was one exception — the Japanese WoZ data on the 0–500ms window when evaluated on the hold/shift task — but otherwise for all languages and genres and all window sizes, our architecture performed better.

6. DISCUSSION

The classic architecture for spoken dialog systems delegates turn taking decisions to a specific module, often the Dialog Manager, which is generally assumed to make these decisions autonomously. This architecture has roots in the traditional view that turn taking follows its own set of rules [30], which govern the flow of dialog, largely involve prosodic signaling, and are largely orthogonal to considerations of semantic content. Although the limitations of this view and this architecture are well-known [31], they are still appealing and have motivated many investigations of turn-taking as a self-contained problem.

If, however, our aim is to build highly responsive dialog systems, it may be time to reconsider this strategy. Recent findings, including our findings above, suggest that, even with models that perform on a par with humans, prediction of the turn status a second in the future is only modestly better than

	% reduction				
Prediction Window	American English	Japanese	Mandarin	Spanish	Canadian French
0 to 250ms	46%	44%	53%	51%	51%
0 to 500ms	28%	23%	36%	34%	35%
0 to 1s	14%	3%	22%	20%	19%
0 to 2s	4%	-5%	9%	7%	7%
0 to 3s	0%	-8%	7%	2%	5%
baseline	0.39	0.34	0.45	0.41	0.43

Table 4. Mean Absolute Error — Telephone Corpora

chance, at best. This suggests that the best strategy for improving the turn taking of dialog systems is no longer to pursue further improvements in turn-taking models, but rather to reduce system latency to make the turn-taking problem easier. That is, the most effective way to improve perceived turn-taking quality is very likely to make dialog systems more incremental. For example, if there is only 250 ms delay between the time that speech input is heard and appropriate action is taken, turn-taking prediction can be very accurate: as seen above, when very recent information is available, turn-taking is much easier.

Prediction Window	MAE	% reduction
0 to 250ms	0.07	76%
0 to 500ms	0.13	55%
0 to 1s	0.19	34%
0 to 2s	0.24	17%
0 to 3s	0.27	9%
baseline	0.29	

Table 5. Results (MAE) — Japanese WoZ

This is not to say, however, that there is no need for further research in turn taking. In multimodal scenarios there are additional sources of information available, such as gaze, gesture, and breathing patterns [32, 33, 34, 14], which are not yet fully exploited. More generally, although many recent models achieve human-level performance in turn-taking, there is no reason why they could not do even better. For example, techniques for rapid adaptation, or joint adaptation, to the turn-taking styles of specific individuals [35, 36], when integrated into LSTM models, may enable us to far exceed human performance. Abandoning the classic assumption of the autonomy of turn taking, to include not only prosodic and discourse marker factors but also higher-level semantic and pragmatic considerations in the decisions [6, 31, 37], also has great promise. Another important need is for techniques for adapatation. Turn-taking patterns and practices in the wild are highly diverse [8] and many types of decisions are involved [9], so we need methods that enable existing turn-taking models to be adapted to a new domain or style of interaction, without requiring the usual costly collection of a new multi-hour corpus to train a brand new model [7, 16].

7. SUMMARY

We have presented a model of turn taking that matches and sometimes exceeds the state labels. We have made the code for this model freely available for general use. Further, using this model to investigate turn-taking across languages and dialog genres, we find large differences in predictability.

8. REFERENCES

[1] Gabriel Skantze, “Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks,” in *SIGdial*, 2017.

[2] Shammur Absar Chowdhury, Evgeny A. Stepanov, and Giuseppe Riccardi, “Predicting user satisfaction from turn-taking in spoken conversations,” in *Interspeech*, 2016, pp. 2910–2914.

[3] Antoine Raux and Maxine Eskenazi, “Optimizing the turn-taking behavior of task-oriented spoken dialog systems,” *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 9, pp. 1–23, 2012.

[4] Jon Gratch, Ning Wang, A. Okhmatovskaia, F. Lamothe, M. Morales, R. J. van der Werf, and Louis-Philippe Morency, “Can virtual humans be more engaging than real ones?,” in *Int’l. Conf. on Human-Computer Interaction; Lecture Notes in Computer Science, 4552*, pp. 286–297. Springer, 2007.

[5] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan

	after a 250ms pause		after a 500ms pause	
Instances	3,022		2,064	
% Hold	84.2%		80.9%	
Model	Replica	Ours	Replica	Ours
Shift: Precision	0.645	**0.672**	**0.740**	0.656
Shift: Recall	0.524	**0.530**	0.406	**0.538**
Shift: F-measure	0.578	**0.593**	0.524	**0.591**
Hold Precision	0.913	**0.915**	0.873	**0.895**
Hold: Recall	0.946	**0.951**	**0.966**	0.933
Hold: F-measure	0.929	**0.933**	**0.917**	0.914

Table 6. Hold/Shift Results — Japanese WoZ Data

Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency, "SimSensei Kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of AAMAS*, 2014, pp. 1061–1068.

[6] Maike Paetzel, Ramesh Manuvinakurike, and David DeVault, "So, which one is it? the effect of alternative incremental architectures in a high-performance game-playing agent," in *SIGdial*, 2015.

[7] Nigel G. Ward and David DeVault, "Challenges in building highly-interactive dialog systems," *AI Magazine*, vol. 37, no. 4, pp. 7–18, 2016.

[8] Nigel G. Ward, *Prosodic Pattterns in English Conversation*, Cambridge University Press, 2019, to appear.

[9] Hatim Khouzaimi, Romain Laroche, and Fabrice Lefevre, "Turn-taking phenomena in incremental dialogue systems," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1890–1895.

[10] Sepp Hochreiter and Jurgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[11] Matt Shannon, Gabor Simko, Shuo-Yiin Chang, and Carolina Parada, "Improved end-of-query detection for streaming speech recognition," *Proc. Interspeech 2017*, pp. 1909–1913, 2017.

[12] Angelika Maier, Julian Hough, and David Schlangen, "Towards deep end-of-turn prediction for situated spoken dialogue systems," in *Interspeech*, 2017.

[13] Matthew Roddy, Gabriel Skantze, and Naomi Harte, "Investigating speech features for continuous turn-taking prediction using lstms," in *Interspeech*, 2018.

[14] Matthew Roddy, Gabriel Skantze, and Naomi Harte, "Multimodal continuous turn-taking prediction using multiscale rnns," in *International Conference on Multimodal Interaction (ICMI)*, 2018.

[15] Kohei Hara, Koji Onoue, Katsuya Takanashi, and Tatsuya Kawahara, "Prediction of turn-taking using multitask learning with prediction of backchannels and fillers," in *Interspeech*, 2018.

[16] Divesh Lala, Koji Inoue, and Tatsuya Kawahara, "Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios," in *International Conference on Multimodal Interaction*, 2018.

[17] Nigel G. Ward, Olac Fuentes, and Alejandro Vega, "Dialog prediction for a general model of turn-taking," in *Interspeech*, 2010.

[18] David Schlangen and Gabriel Skantze, "A general, abstract model of incremental dialogue processing," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 2009, pp. 710–718.

[19] Timo Baumann, *Incremental Spoken Dialogue Processing: Architecture and Lower-level Components*, Ph.D. thesis, Universität Bielefeld, Germany, 2013.

[20] Mike Brooks, "Voicebox: Speech processing toolbox for Matlab," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 2018.

[21] Nigel G. Ward, "Midlevel prosodic features toolkit," https://github.com/nigelgward/midlevel, 2017.

[22] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[23] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[24] Diego Aguirre, "LSTM turn-taking code," https://github.com/aguirrediego/LSTM-Language-Model, 2018.

[25] Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al., "The HCRC map task corpus," *Language and Speech*, vol. 34, pp. 351–366, 1991.

[26] Tanya Stivers, Nicholas J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al., "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10587–10592, 2009.

[27] Pablo Brusco, Juan Manuel Perez, and Agustin Gravano, "Cross-linguistic study of the production of turn-taking cues in American English and Argentine Spanish," *Interspeech*, pp. 2351–2355, 2017.

[28] Yasuo Horiuchi, Yukiko Nakano, Hanae Koiso, Masato Ishizaki, Hiroyuki Suzuki, Michio Okada, Makiko Naka, Syun Tutiya, Akira Ichikawa, Horiuchi Yasuo, et al., "The design and statistical characterization of the Japanese map task dialogue corpus," *Journal of the Japanese Society for Artificial Intelligence*, vol. 14, pp. 261–272, 1999.

[29] Nigel G. Ward and Wataru Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, pp. 1177–1207, 2000.

[30] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, pp. 696–735, 1974.

[31] Ethan O. Selfridge and Peter A. Heeman, "Importance-driven turn-bidding for spoken dialogue systems," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 177–185.

[32] Sean Andrist, Michael Gleicher, and Bilge Mutlu, "Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 2571–2582.

[33] Marcin Wlodarczak and Mattias Heldner, "Respiratory belts and whistles: A preliminary study of breathing acoustics for turn-taking," in *Interspeech*, 2016.

[34] Margaret Zellers, David House, and Simon Alexanderson, "Prosody and hand gesture at turn boundaries in Swedish," in *Speech Prosody*, 2016.

[35] Rivka Levitan, Stefan Benus, Agustin Gravano, and Julia Hirschberg, "Entrainment and turn-taking in human-human dialogue," in *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction*, 2015, pp. 44–51.

[36] John Grothendieck, Allen L. Gorin, and Nash M. Borges, "Social correlates of turn-taking style," *Computer Speech and Language*, vol. 25, pp. 789–801, 2011.

[37] Andrea L Thomaz and Crystal Chao, "Turn-taking based on information flow for fluent human-robot interaction," *AI Magazine*, vol. 32, no. 4, pp. 53–63, 2011.