INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

An autoregressive conversational dynamics model for dialogue systems

Matthew McNeill[1], *Rivka Levitan*[2]

[1]CUNY Graduate Center, New York, NY, USA
[2]Brooklyn College, Brooklyn, NY, USA

mmcneill@gradcenter.cuny.edu, rlevitan@brooklyn.cuny.edu

Abstract

Conversational partners adapt their speech to one another in a phenomenon called *entrainment*. While entrainment behaviors are associated with a variety of positive conversational outcomes, they are rarely implemented in dialogue systems due to their poorly understood mechanics. Conversational dynamics models could discover entrainment behavior in a dialogue corpus, but to date they have not been designed for or evaluated in dialogue systems. In this paper, we propose an autoregressive model specifically for use in a dialogue system. We evaluate its ability to predict features for upcoming conversational turns, and show it outperforms several baseline models. Additionally, we analyze its attention mechanism to explain which turns it finds useful for predicting upcoming speech features. Finally, we discuss its potential for future deployment in a live dialogue system.

Index Terms: dialogue systems, entrainment, conversational dynamics

1. Introduction

When people converse, they adjust their speaking patterns in response to one another in a complex set of behaviors variously referred to as *entrainment*, *adaptation*, or *alignment*. Entrainment occurs to some extent in nearly every conversation, but it has not been widely implemented in spoken dialogue systems. Dialogue agents usually speak with predetermined prosody regardless of their partner's speech, potentially leaving a significant amount of user goodwill on the table. Previous studies found that implementing specific entrainment behaviors in dialogue systems improves rapport [1], trust [2, 3], and learning [4], and that users generally prefer entraining agents to non-entraining agents [5]. However, the mechanics of entrainment are poorly understood, and attempts to explain them are difficult to reproduce [6]. This suggests that the complex entrainment behaviors in human-human conversation would be better modeled by general-purpose conversational dynamics models. Recently, several promising models were introduced and evaluated in terms of their ability to predict prosodic qualities of upcoming conversational turns. However, none were specifically designed for dialogue systems, and none have been evaluated conversing with human partners. Aside from overcoming the technical challenges of adapting these models for a live setting, it is unclear how they would perform.

In this paper, we present an autoregressive conversational dynamics model to direct the output of controllable text-to-speech (TTS) in a dialogue system. Like other recent work, our model predicts acoustic and prosodic speech features of upcoming conversational turns by maintaining and attending to a history of prior turns. Our contributions are as follows:

1. To mimic human and agent roles in a dialogue system, we assign them to conversational partners in our corpus. Our model only predicts upcoming speech features for the agent.
2. Our model is trained and evaluated with a partially autoregressive process that mimics the setting of a dialogue system.
3. Our model outputs speech features associated with entrainment behavior, including pitch, intensity, jitter, shimmer, noise-to-harmonics ratio, and speaking rate.
4. We analyze our model's attention mechanism to explain how it determines the most relevant historical turns for predicting upcoming speech features.

We begin with a review of the state of user-responsive dialogue systems in Section 2. In Section 3, we describe the setting of a dialogue system, the architecture of our model, and our feature extraction process. In Section 4, we establish a baseline and perform several comparative experiments with the Fisher corpus. In Section 5, we discuss our results and present a novel analysis of our model's attention mechanism to demonstrate that it identifies patterns of behavior in the corpus. Finally, in Section 6, we discuss future plans for the model.

2. Related Work

Most attempts to introduce entrainment behavior in dialogue systems approximate *proximity* as defined by [7]. This is accomplished by the simple mimicking of selected speech features extracted from recordings of a human conversational partner. For example, a tutor dialogue agent in [1] employs a variety of methods to manipulate a TTS pitch contour in response to its partner's speech, including directly adopting their pitch contour and shifting a generated contour to match their pitch. A game requiring players to seek advice from a conversational avatar in [5] matches its partner on the intensity and speech rate of their previous conversational turn. A similar advice game in [3] adjusts its pitch, intensity, and speaking rate according to the player's most recent turn relative to a baseline obtained at the beginning of the game. Finally, a chit-chat system in [2] matches its partner on speech rate, pitch, and loudness, and includes lexical entrainment by mimicking its partner's pronoun use, repetition of terms, and utterance length.

In contrast to implementations of specific entrainment behaviors, conversational dynamics models attempt to discover speaking patterns in a dialogue corpus. [8, 9], on which this work is partially based, maintains a window of the past 1, 5, or 10 timesteps of a conversation's history and uses this window to predict the upcoming turn's energy, pitch range, and speaking rate. [10] combines a language model and a prosody model, using a transcribed conversational prompt and word-level pitch values to jointly predict a textual response with word-level pitch

10.21437/Interspeech.2023-2525

differentials. Finally, [11] uses an entire linguistic and prosodic dialogue history in addition to the text of an upcoming utterance as input to a TTS, demonstrating that access to dialogue context improves the realism of synthesized speech in a conversational setting.

3. Method

3.1. Setting

We adapt the spoken dialogue system architecture from [5] as a setting for our model. A typical dialogue system transcribes and segments user speech with automatic speech recognition (ASR), generates a textual response, and converts the response to audio with TTS. An enhanced dialogue system incorporating our model extends this process. It extracts acoustic-prosodic features from recorded speech and combines them with ASR transcriptions as input to our model, which encodes them and appends the result to an in-memory dialogue history. Using the history, it predicts appropriate acoustic-prosodic features for the response. The predicted features and textual response are passed to a controllable TTS for speech generation, and are themselves encoded and appended to the dialogue history.

The features we can predict depend on our choice of TTS. Commercial systems such as Cepstral, Alexa, and Google TTS offer controls for pitch, intensity, and speaking rate. Controllable variants of neural TTS like Tacotron [12] can be trained with features extracted from a speech corpus for utterance-level control [13]. For this work, we assume our dialogue system has adapted this variant of Tacotron trained with controls for the features described in Section 3.2.

3.2. Data

We train and evaluate our model with the Fisher corpus [14], a large collection of dyadic telephone conversations that have been transcribed and segmented. Because there is no visual component to the conversations, any entrainment behavior can only occur over lexical, acoustic, and prosodic dimensions. We consider the corpus an approximation of conversations that might occur in a general-purpose chit-chat dialogue system.

Transcripts were tokenized with Torchtext and converted to sequences of 50-dimensional GloVe word embeddings [15]. Features were extracted from speech segments with Praat [16], discarding turns too short for processing. The resulting dataset contains 11,699 conversations with 3,264,021 turns. We chose 7 features based on their association with entrainment behavior [7] and their use in controllable TTS [13]:

- Mean log-pitch, or f_0.
- Log-pitch range, or the difference between the 95th- and 5th-percentile log-pitch.
- The mean intensity of voiced frames.
- Jitter, shimmer, and noise-to-harmonics ratio (NHR) from voiced frames.
- Speaking rate, or the mean duration of syllables in the turn.

Features were normalized per-speaker and per-conversation. As in [13], we normalize by finding the median (M) and standard deviation (σ) of each feature, then project values in the range $[M - 3\sigma, M + 3\sigma]$ to $[-1, 1]$. We omit the step of clipping feature values to $[-1, 1]$ because removing values exceeding typical speech ranges could degrade model performance, and clipping model output can be performed prior to engaging the controllable TTS.

To mimic the environment of a live dialogue system, we designate one speaker in each conversation as the agent and their partner as the human. For consistency, the first to speak is designated as human. Turns were annotated with a one-hot vector indicating the speaker's assumed role.

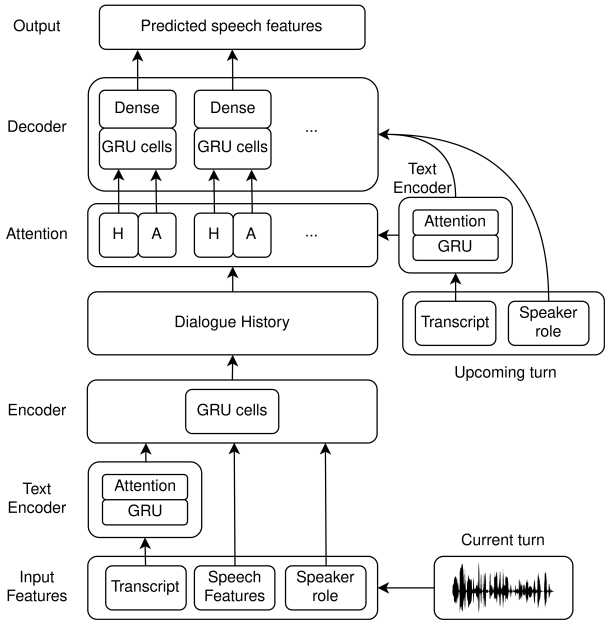

Figure 1: *Diagram of our model predicting speech features of an upcoming conversational turn.*

3.3. Baseline model

As a baseline, we adopt a fixed-window recurrent model similar to [8, 9]. We use a bidirectional 2-layer GRU to encode a representation vector of 10 previous turns, where the next to be predicted is an agent turn. The final GRU output is used to predict features by passing it through 7 separate 2-layer dense networks with ELU activation, one for each output feature. As part of our evaluation, we vary the decoder inputs and contents of the 10-turn representation vector as described in Section 4.1.

3.4. Our model

Our model is depicted in Figure 1. It consists of an encoder and 7 decoders, one for each output feature. The current turn's transcript is encoded with a 2-layer bidirectional GRU with additive attention, and concatenated with its extracted speech features and speaker role vector. These inputs are encoded to a hidden representation vector with a 2-layer GRU cell and appended to an accumulated dialogue history.

Like [17], each decoder is equipped with a pair of attention mechanisms: one that attends to historical agent turns (labeled A in Figure 1), and one that attends to historical human turns (labeled H). Both attention mechanisms are given the decoder hidden state and the encoded transcript of the upcoming turn as context. The outputs from both attention mechanisms are concatenated and decoded to an output speech feature with 2 GRU cells and a 2-layer dense network with ELU activation. The encoded transcript and speaker role of the upcoming turn are given as context. All decoder outputs are concatenated into a new feature vector containing the 7 features expected to be found in the upcoming turn.

4. Experiments

The corpus was split into a 90% training and 10% test set and used to evaluate 4 variants of the baseline model against our model. To test each model's ability to discover consistent patterns in a variety of conversations and speaking styles, the training set was used to conduct a 5×2 cross-validation. This consists of 5 2-fold cross-validations, with significance determined by Alpaydin's F-test [18]. All models use AdamW optimization. The baseline variants were trained with a learning rate of 1^{-4} and a batch size of 64, and our model with a learning rate of 5^{-4} and a batch size of 32. All models were allowed to train until validation error did not improve for 25 epochs. Checkpoints from the best-performing epoch were used for evaluation.

The model was written in PyTorch and PyTorch Lightning. Preprocessing code, models, and hyperparameters are available on GitHub[1]. Cross-validation folds were trained on a combination of NVIDIA 3090 and 2080 Ti GPUs. Each fold of our model took approximately 2 days to complete, and each fold of our baselines took between 7 hours to 1 day to complete.

4.1. Baseline model

We train 3 variants of the baseline model with an increasing amount of features in the dialogue history, plus a 4th variant with text from the upcoming turn as additional decoder input:

1. **Features (F)**: The history contains speech features.
2. **F + embeddings (E)**: The history contains speech features and GloVe embeddings converted to a hidden vector by a 2-layer bidirectional GRU encoder with additive attention.
3. **F + E + speaker role (SR)**: The history contains speech features, encoded GloVe vectors, and speaker role.
4. **F + E + SR with embedding decoder input**: The history contains speech features, encoded GloVe vectors, and speaker role. Additionally, GloVe vectors encoded from the upcoming turn are concatenated with the encoded dialogue history prior to decoding.

Like [8], variants 1-3 are intended to establish that additional information in the dialogue history reduces error in predicting speech features for the upcoming turn. The 4th variant brings the baseline as close as possible to our model without fundamentally altering its structure, allowing us to demonstrate the effectiveness of our model's attention mechanisms, complete dialogue history, and training procedure.

For each minibatch, we randomly choose a turn n from each conversation to predict, and assemble a window of prior turns $n-11$ to $n-1$. In cases where this goes past the beginning of the conversation, the window is left-padded with zeros. For backpropagation, only agent turns are used to compute MSE loss against ground-truth feature values.

To evaluate the baseline variants, we approximate what would be required to use them in a dialogue system. We generate consecutive windows $n-11$ to $n-1$ predicting n, starting where n is the first agent turn and sliding forward until we reach the end of the conversation. To evaluate performance differences between autoregressive and non-autoregressive inference, we perform two evaluations: one replicating the testing procedure in [8] where ground-truth values are always present in the window, and one where agent turn predictions are autoregressively fed back into the sliding window.

[1] https://github.com/mattm458/conv-dynamics-dialogue-systems

4.2. Our model

Our model uses a partially autoregressive training procedure designed to mimic the environment of a live dialogue system, where the system controls its own output in response to a human partner. Training consists of encoding and decoding each conversational turn in sequence to predict the speech features of the upcoming turn. For upcoming human turns, predicted speech features are discarded, and ground-truth extracted feature values are given as input with teacher forcing. For upcoming agent turns, predicted speech features are fed back into the encoder. The same process is used for inference. During backpropagation, only predicted agent turns are used to compute the MSE loss against ground-truth feature values.

4.3. Attention analysis and out-of-domain evaluation

If our model successfully discovers speaking patters in the corpus, attention scores should be similar across training runs. If it fails to do so, attention scores should be different across training runs. We evaluated pairs of folds with partially overlapping training data against the test set, extracted attention scores, and determined the extent to which top 10 and top 5 scores overlap.

Additionally, we performed an out-of-domain evaluation against a subset of the NXT-format Switchboard corpus [19]. Though Switchboard is similar to Fisher in both structure and content, we consider it out-of-domain due to the corpuses' partially non-overlapping conversation topics, different turn segmentation, and different participant demographics. These differences are analogous to using our model in a live dialogue system, where factors like turn segmentation, participants, and conversation topics are not the same as our dataset.

This final evaluation with NXT-format Switchboard gave us an opportunity to explain our model's attention scores. Unlike the Fisher corpus, NXT-format Switchboard is partially annotated for dialogue acts, which we incorporate into our analysis. We constructed a linear model to predict z-score normalized attention scores at each historical timestep from the position of the score in the dialogue history as a percentage of its length, the distance between the historical turn's speech feature values and predicted feature values, and whether the dialogue act of the upcoming turn matches the dialogue act of the historical turn. Though our model does not use them as inputs, the attention mechanism has access to historical and upcoming lexical content in encoded GloVe embeddings. Because dialogue act annotations describe a turn's purpose and act as an approximation of its content, we hypothesize that historical turns annotated with the same dialogue act as the upcoming turn will be more useful than historical turns with a different dialogue act. Dialogue acts were consolidated into 5 categories [20]: Statement+Opinion, Question, Backchannel, Answer+Agree, and Other, and multiple dialogue acts in a turn were resolved with majority voting.

5. Results

5.1. Cross-validation results

A comparison between autoregressive and non-autoregressive inference in the baseline variants indicated slightly worse performance for autoregressive inference. Because it is more representative of conditions in a live dialogue system, we only report results from autoregressive inference.

Results are shown in Table 1. We report the smooth L1 loss of agent turn predicted features compared with ground-truth values. Like [8], additional features in the baseline window

Table 1: *5×2 cross-validation smooth L1 loss for each model autoregressively predicting speech features for agent turns. Results in* ***bold*** *are significantly better than the previous row ($p < 0.05$). Results marked with * are highly significant ($p < 0.001$).*

	Enc.	Dec.	All	f_0	f_0 Range	Intensity	Jitter	Shimmer	NHR	Rate
Bsln.	F		0.0541	0.0529	0.0569	0.0537	0.0537	0.0546	0.0544	0.0528
	F+E		**0.0538***	**0.0525**	**0.0568**	**0.0528***	0.0537	**0.0545**	0.0545	**0.0521**
	F+E+SR		**0.0533***	**0.0518**	**0.0567**	0.0525	0.0536	**0.0544**	0.0542	**0.0495**
	F+E+SR	E	**0.0464***	**0.0507***	**0.0547***	**0.0453***	**0.0512***	**0.0525***	**0.0513***	**0.0192***
Ours	F+E+SR	E+SR	**0.0442***	**0.0486***	**0.0536***	**0.0421***	**0.0497***	**0.0511***	**0.0493***	**0.0152***

Table 2: *Percent of top-scoring attention scores overlapping across cross-validation folds.*

Attention	Feature	Overlap (10)	Overlap (5)
Agent	f_0	27.84	18.19
	f_0 Range	30.40	20.33
	Intensity	28.25	18.43
	Jitter	20.24	13.35
	Shimmer	17.37	12.48
	NHR	16.94	11.75
	Rate	36.07	27.19
Human	f_0	63.58	49.57
	f_0 Range	34.71	25.33
	Intensity	64.12	50.78
	Jitter	21.49	19.37
	Shimmer	21.49	16.67
	NHR	28.06	24.36
	Rate	40.76	43.61

improved performance. The addition of speaker role and encoded GloVe vectors resulted in highly significant ($p < 0.001$) reduction in overall error, with significant ($p < 0.05$) improvements for most features individually. These results suggest that, while a combination of prosodic and lexical features from both speakers are necessary to predict upcoming features, the model considers each speaker differently. However, the biggest reduction in error came when introducing encoded GloVe vectors of the upcoming turn, suggesting that the speaker's intent plays a significant role in predicting speech features.

Our model exhibited a highly significant reduction in error across all features, outperforming the baseline. We believe there are three reasons for this: first, using the entire dialogue history gives the decoder more prosodic and lexical context; second, its attention mechanism gives it the ability to ignore irrelevant turns; and third, both the encoder and decoder are following every step of the conversation, even if the decoder's output for human turns are discarded. Evaluating the best-performing fold of our model with the test set resulted in an overall error of 0.0442. This is within error rates seen in cross-validation, and demonstrates the ability of our model to generalize.

5.2. Attention analysis and out-of-domain evaluation

A summary of attention score similarity across folds is shown in Table 2. Human attention mechanisms attended to the same turns more frequently than agent attention mechanisms. Pitch, intensity and speaking rate overlapped the most in the top 10 and top 5 attention scores. We note that the most consistently selected human turn features correspond with the lowest prediction error: pitch, intensity, and speaking rate.

Evaluating the best-performing fold of our model with Switchboard data resulted in an overall error of 0.0498. This is not far outside the error rates seen in cross-validation, and shows that our model can generalize to unfamiliar data.

Our linear models were most successful at explaining the variance in attention scores when our model predicts mean pitch ($R^2 = 0.373$) and intensity ($R^2 = 0.353$) from human turns, and pitch range ($R^2 = 0.324$) from agent turns. R^2 values were generally lower for attention scores from agent turns, consistent with the interpretation that it is more difficult to pinpoint the most relevant turns in the agent's history.

When predicting pitch from human turns, high attention scores are associated with recent turns, matching dialogue acts, and similar feature values except pitch and speaking rate. For intensity, high attention scores are associated with recent turns, matching dialogue acts, and similar feature values except speaking rate. In contrast, when predicting pitch range from agent turns, high attention scores are associated with mismatched dialogue acts, earlier turns, dissimilar pitch, jitter, shimmer, and speaking rate; and similar pitch range, intensity, and NHR. A full listing of results is available online.[2]

6. Conclusions and Future Work

In this paper, we introduced an autoregressive recurrent conversational dynamics model to predict a dialogue agent's speech features and direct a controllable TTS. We described a training and evaluation procedure to mimic a dialogue system, and used it to train and evaluate our model against a baseline. Our model outperformed the baseline, and demonstrated comparable results in an out-of-domain speech corpus. We evaluated our model's attention mechanism and showed that, for some predicted features, we can partly explain the characteristics of historical turns it is likely to attend to, yielding novel insights into how adaptation to one's interlocutor can occur in a manner that is not necessarily linear or tied to adjacency.

There are several avenues for future work with our model. First, this work does not consider the TTS component of a dialogue system. We hope to determine the extent to which neural TTS can synthesize speech in accordance with our predicted feature values, and improve output in cases where the TTS unable to satisfactorily achieve our targets. Additionally, we hope to expand our model with the rich conversation data collected in recent corpora like [21], including speaker demographics and turn-level emotion annotation. Our ultimate goal is to evaluate our model in a live dialogue system with human partners, comparing it against existing matching entrainment strategies.

[2] http://www.sci.brooklyn.cuny.edu/~levitan/speechlab/conv-dynamics-dialogue-systems/

7. References

[1] N. Lubold, H. Pon-Barry, and E. Walker, "Naturalness and rapport in a pitch adaptive learning companion," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 103–110.

[2] R. Hoegen, D. Aneja, D. McDuff, and M. Czerwinski, "An End-to-End Conversational Style Matching Agent," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, ser. IVA '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 111–118. [Online]. Available: https://doi.org/10.1145/3308532.3329473

[3] S. Benus, M. Trnka, E. Kuric, L. Marták, A. Gravano, J. Hirschberg, and R. Levitan, "Prosodic entrainment and trust in human-computer interaction," in *Proc. 9th International Conference on Speech Prosody 2018*, 2018, pp. 220–224. [Online]. Available: http://dx.doi.org/10.21437/SpeechProsody.2018-45

[4] J. Thomason, H. V. Nguyen, and D. Litman, "Prosodic Entrainment and Tutoring Dialogue Success," in *Artificial Intelligence in Education*, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 750–753.

[5] R. Levitan, Štefan Beňuš, R. H. Gálvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg, "Implementing Acoustic-Prosodic Entrainment in a Conversational Avatar," in *Proc. Interspeech 2016*, 2016, pp. 1166–1170.

[6] A. Weise and R. Levitan, "Investigating the influence of personality on acoustic-prosodic entrainment," in *Proc. Interspeech 2022*, 2022, pp. 3093–3097.

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

[8] S. Fuscone, B. Favre, and L. Prévot, "Neural Representations of Dialogical History for Improving Upcoming Turn Acoustic Parameters Prediction," in *Proc. Interspeech 2020*, 2020, pp. 4203–4207.

[9] S. Fuscone, B. Favre, and L. Prevot, "Comparing monological and dialogical neural representations of dialogue history for predicting the acoustic parameters of an upcoming conversational turn," *Lingue e linguaggio*, vol. 20, no. 2, pp. 259–288, 2021.

[10] Y. Yamazaki, Y. Chiba, T. Nose, and A. Ito, "Neural Spoken-Response Generation Using Prosodic and Linguistic Context for Conversational Systems," in *Proc. Interspeech 2021*, 2021, pp. 246–250.

[11] Y. Nishimura, Y. Saito, S. Takamichi, K. Tachibana, and H. Saruwatari, "Acoustic Modeling for End-to-End Empathetic Dialogue Speech Synthesis Using Linguistic and Prosodic Contexts of Dialogue History," in *Proc. Interspeech 2022*, 2022, pp. 3373–3377.

[12] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[13] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable Neural Text-to-Speech Synthesis Using Intuitive Prosodic Features," in *Proc. Interspeech 2020*, 2020, pp. 4432–4436.

[14] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2004/pdf/767.pdf

[15] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[16] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[17] R. Lahiri, M. Nasir, C. Lord, S. H. Kim, and S. Narayanan, "A Context-Aware Computational Approach for Measuring Vocal Entrainment in Dyadic Conversations," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[18] E. Alpaydin, "Combined 5 x 2 cv F test for comparing supervised classification learning algorithms." *Neural computation*, vol. 11, pp. 1885–92, Nov 1999.

[19] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language resources and evaluation*, vol. 44, pp. 387–419, 2010.

[20] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 09 2000. [Online]. Available: https://doi.org/10.1162/089120100561737

[21] A. Weise, M. McNeill, and R. Levitan, "The Brooklyn Multi-Interaction Corpus for Analyzing Variation in Entrainment Behavior," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 1721–1731. [Online]. Available: https://aclanthology.org/2022.lrec-1.183