Interspeech 2024
1-5 September 2024, Kos, Greece

"Well", what can you do with messy data? Exploring the prosody and pragmatic function of the discourse marker "well" with found data and speech synthesis

Johannah O'Mahony[1], *Catherine Lai*[1], *Éva Székely*[2]

[1]University of Edinburgh, UK [2]KTH Royal Institute of Technology, Stockholm, Sweden

johannah.o'mahony@ed.ac.uk

Abstract

Recently, there has been growing interest in the synthesis of conversational speech prosody. Conversational prosody is variable and carries many pragmatic functions. As speech synthesis research moves to using large amounts of untranscribed data, it is crucial that we understand the subtle pragmatic differences prosody can make. This study focuses on discourse markers, which are linguistic elements that perform various communicative functions, with their specific roles often linked to their prosodic realisation. In this paper, we explore the prosodic realisation of *well* using an unlabelled corpus of conversational speech. We use clustering to explore the variation in its prosodic realisation and identify common patterns in a data-driven manner. We synthesise the cluster centroids using controllable speech synthesis. Finally, we evaluate how the prosodic realisation of *well* affects the meaning of an utterance.

Index Terms: conversational speech synthesis, pragmatics, prosody

1. Introduction

Since the improvement in the synthesis of read speech in Text-to-Speech synthesis (TTS), there has been growing interest in synthesising conversational speech, in particular in synthesising expressive conversational prosody. Conversational prosody is highly variable and context-dependent, and it has many functions in conversation, such as signalling salient information, managing turn-taking, and expressing stance/attitude. To model natural conversational prosody, recent approaches in TTS research are beginning to use large amounts of unlabelled *found data* [1, 2], which have not been created for the purpose of building TTS models [3].

One of the benefits of using found data is that we can use larger amounts of data to capture a broader range of prosodic variation and spontaneous/conversational speech phenomena such as backchannels [4], filled pauses [5, 6] and discourse markers [7, 8]. However, this approach also presents a significant challenge, namely accounting for the variation in the data. Without accounting for variation in the data, we risk synthesising *average prosody* [9] or an inappropriate realisation given the dialogue context. Synthesising a realisation unsuitable for its pragmatic function could have negative effects on spoken interactions. So, our goal is to identify prosodic patterns in found data, and to explore the pragmatic functions of prosody in order to aid the synthesis of pragmatically felicitous prosody.

In this work, we explore how we can begin to understand the variation in an unlabelled dataset of conversational speech. Here, we focus on discourse markers (DMs), which have received relatively little attention in speech synthesis research, but which make up a significant subset of the top twenty most frequent words in conversational English due to their multifunctionality in conversation [8]. From a prosody perspective, DMs are an interesting case of the pragmatics-prosody interface as their function can be related to their prosodic realisation [10]. In this study, we explore the variation in prosodic renditions of *well*, one of the most studied discourse markers in English [11].

The use of *well* carries many functions including signalling a topic shift [12, 13], expressing stance [13], signalling that an answer is insufficient or dispreferred [13], or hedging a disagreement [14, 11]. The prosodic realisation has also been found to differ for *well* depending on its function, for example based on the presence of pauses [15], duration [16] and pitch realisation [10]. Most studies on the DM *well* rely on human-annotated labels, however our proposed method explores pragmatic meaning in an entirely data-driven manner using unlabelled found data. In this study we:

1. develop a method for extracting a representative corpus of discourse markers from realistic speech data, for the example of *well*.
2. explore the prosodic realisation of *well* in a data-driven manner using clustering.
3. synthesise prosodic variation on the discourse marker *well* based on the cluster centroids using controllable TTS.
4. evaluate the effect of the prosodic realisation of *well* on the level of *agreement/disagreement* perceived by a listener.

2. Method

2.1. Curating a Discourse Marker Corpus

2.1.1. Conversational Data

We use the CANDOR corpus [17] of 1656 2-channel open-domain conversations of $25\,\mathrm{min}$ recorded on a video call between two strangers. This corpus has been automatically transcribed using Amazon Web Services (AWS) Automatic Speech Recognition (ASR) and therefore contains no ground-truth transcription or additional annotations. The corpus contains a wide range of recording conditions, background noise, and occasional channel leakage if speakers were not using headphones.

2.1.2. Data Filtering

We split the CANDOR corpus into interpausal-units (IPUs) based on a silence threshold of $180\,\mathrm{ms}$. Common one word backchannels were removed and the IPUs were classified following the communicative state classification in [18]. We then extracted each IPU if the IPU-initial unigram was the token *well*. We classified all IPUs depending on their position in a turn and kept IPUs which were turn-initial, turn-medial, or an overlap initiated by the speaker.

10.21437/Interspeech.2024-2122

Due to the varying audio and transcription quality, we ran a number of filtering steps to improve the quality of the dataset. We first used the ASR model Parakeet[1] to transcribe each IPU in the dataset. We then compared the new transcription with the original ASR transcript and kept IPUs if both agreed on the presence of the initial *well* token. To estimate the speech-to-noise ratio (SNR) of each recording, we used the Brouhaha model [19] and removed IPUs with an SNR ratio of 25 or lower.

Each IPU was then aligned using the Montreal Forced Aligner (MFA) [20] and IPUs where forced alignment failed were removed. Due to the nature of the data and the fact that alignments for spontaneous speech may contain more errors than read speech [21], we spliced out the word of interest *well* and ran the spliced audio through Parakeet ASR in a second pass to verify that the spliced word was correct and didn't contain additional words. Finally, we used the presence of silence after the *well* token as a proxy for a prosodic boundary, as in this study we are interested in tokens which occur in their own prosodic phrase. Details of IPUs remaining after each filtering step and final data quantities can be found in Table 1. While a significant amount of the potential data is lost, a trade-off has to be considered between quantity and quality when using found data and we chose a conservative approach. In total, there are 484 speakers in the *well* dataset.

Table 1: *Remaining IPU quantity after each filtering step*

Filter	Number of IPUs
Raw	8095
Parakeet Errors	8067
Parakeet match to ASR	5780
SNR	5319
MFA Errors	5262
Parakeet on spliced *well*	1883
Post-*well* silence	942

2.1.3. Acoustic Information

To explore the prosodic realisation of the discourse marker *well*, we extracted a number of features relating to the F_0 contour, as well as the duration of each realisation. Global F_0 information was first extracted to normalise the word-level F_0 features. To do this, we extracted the F_0 per speaker across every conversation that they took part in using Parselmouth [22] with the Praat autocorrelation F_0 extraction function [23].

After splicing the *well*-tokens from each IPU, word-level F_0 was extracted using the Praat filtered autocorrelation function in a two-step manner following [24], with octave jump removal and interpolation over missing F_0 values. F_0 was converted to semitones using the *global* speaker median. F_0 values were then z-score normalised using the *global* speaker mean and standard deviation in semitones. We chose to normalise by global features to capture where the *well* realisation fell in a speaker's global range, which may be pragmatically meaningful, and normalising the F_0 on the utterance-level would not capture this.

To describe the F_0 characteristics of each instance of *well*, we used a third-order Legendre polynomial to model each F_0 contour. Legendre polynomials are a series of orthogonal polynomials and have been used in linguistic research to mathematically describe F_0 contours e.g. [25]. The first three coefficients characterise the F_0 height (LC0), slope (LC1) and curvature (LC2) respectively. To extract the polynomials we time-normalised the F_0 contour between -1 and 1 and fit a third-order Legendre polynomial and extracted the first three coefficients. Finally we extracted the duration of each *well*-token according to the forced alignment. We further removed tokens if their duration fell below the first percentile or above the 99th percentile (15 removed) and one token failed during F_0 tracking.

2.1.4. Clustering Prosodic Features

In this study, we are interested in the variation in the realisation of *well* in our dataset in terms of the F_0 contour and duration. To explore this variation, we used k-means clustering. Clustering has been used in previous research for finding new patterns in data in prosodic research [26, 27]. By partitioning the acoustic space into clusters of similar prosodic features, we can then use the cluster centroid, or average of each cluster, to control a TTS model. For our clustering we used four features: the duration of each token in seconds, F_0 height (LC0), slope (LC1), and curvature (LC2). This led to a feature vector of four dimensions for each of our 926 *well* samples.

To cluster the data, we used the k-means algorithm from the Python package sklearn [28]. We scaled each dimension using sklearn RobustScaler which scales the data based on percentiles and is robust to outliers. This allowed us to retain values at the tails of our distribution in the clustering. We chose to split the feature space into 20 clusters to cover enough variation in the acoustic space for stimuli creation for our experiment. We do not however assume stringent categories of prosodic realisations, but use this as a tool to partition groups of closely related features.

Results of the clustering are visualised using tSNE in Figure 2 and the distribution of each prosodic feature per cluster can be seen in Figure 1. The tSNE visualization reveals that while some clusters, especially those at the center of the feature space, overlap, peripheral clusters are smaller and more distinct. Such overlap is expected given the dataset's diversity. Nonetheless, informal auditory evaluation of the samples closest to the cluster centroids revealed a perceptual similarity among them.[2]

2.2. TTS

2.2.1. TTS Model

In this work we use FastPitch [29], which is a non-autoregressive end-to-end speech synthesis model based on the transformer architecture which predicts a sequence of mel-spectrogram frames from the input phone sequence. The model uses three variance adapters to condition each phone on F_0, intensity and duration. During training, the model is conditioned on the ground-truth values while a predictive model per feature is trained. During inference, the predicted values can be used to condition the model, or specific values can be passed to the model to control the prosody. We provide additional conditioning parameters to our model following [30]. Here phrase-level slope and word-level Legendre polynomial coefficients can be used to condition the model hierarchically to specify the shape of the F_0 contour. Therefore at synthesis time, the user can specify a word-level accent shape by changing the coefficients

[1] https://huggingface.co/nvidia/parakeet-tdt-1.1b

[2] Samples page https://johannahom.github.io/IS-2024/

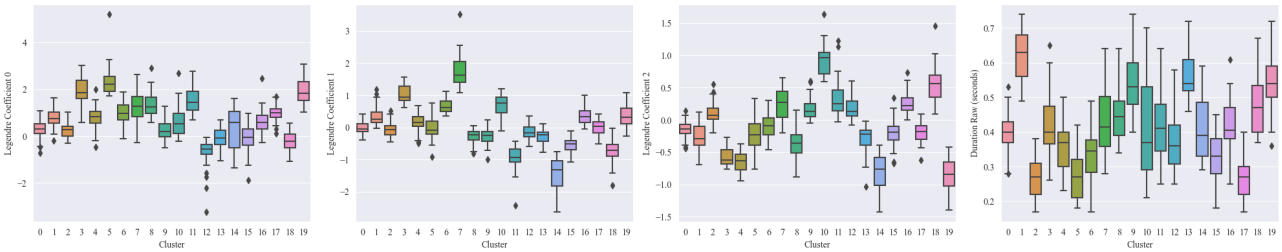

Figure 1: *Boxplots of the features used in k-means for each of the 20 clusters*

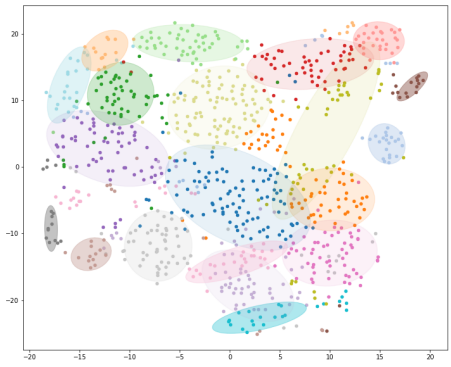

Figure 2: *tSNE plot showing clusters found in k-means.*

of the input data. Specifically, for our purposes, this method enables the utilisation of the Legendre coefficients from the centroids of each cluster along with their durations to generate specific prosodic renditions of *well*.

2.2.2. *Corpora and training*

For TTS training, we use a mix of spontaneous and read speech. Using read speech helps to create a stable speech synthesis voice, while the addition of spontaneous data exposes the model to conversational prosodic patterns. For our read speech, we used the LJSpeech corpus [31] consisting of 13100 utterances of audiobook data read by a female US English speaker.

For our spontaneous speech, we use the AptSpeech corpus, a publicly available multimodal, multi-party corpus of spontaneous conversational speech [32]. It consists of 15 interactions between a moderator and two varying participants playing a collaborative game. The recordings of the moderator have been annotated and segmented into breath groups following [1]. Next to the moderator's spontaneous speech, the moderator recorded additional prompts in read speech style. A summary of the exact data used in training is found in Table 2.

Table 2: *Data used for TTS model building*

Dataset	Train	Val	Test
Male Speaker (Read)	2165	4	27
Male Speaker (Conversational)	5194	21	49
Female Speaker (Read)	12693	25	278
Total	20052	50	354

We trained a multi-speaker TTS model using the corpora described in section 2.2.2. The specific data distribution can be found in Table 2. Our model was trained on a single GPU with a batch size of 16 for 500 epochs. We used the pre-trained NeMo Hifi-GAN vocoder[3] to generate all audio using the LJSpeech speaker label.

3. Evaluation

3.1. Objective

As mentioned in section 1, the use of *well* serves multiple functions in dialogue including hedging, signalling dispreferred answers and agreement/disagreement. In this study we focus on *well* and its use to signal agreement/disagreement. Here we investigate how much weight the prosody of the discourse marker has when the context it precedes remains constant. We present a perceptual study on the use of *well* in signalling agreement or disagreement when followed by the words *yes* (positive polarity) and *no* (negative polarity). Here we ask: Does the level of agreement expressed by the speaker change as a function of the prosody of *well*? To test this, we use the centroids of the clusters found in section 2.1.4 to generate 20 stimuli per polarity type.

3.2. Stimuli Creation

We use the Switchboard NXT corpus [33] consisting of 1126 conversations labelled for Dialogue Act (DA) information. All utterances were extracted which began with *well*. We limited the DA categories to *answers*, *no*, *hedge*, *agree*, *yes*, *reject*, *affirm*, *neg*, *ans-dispreferred* and *maybe* and only chose utterances whose exact textual content appeared two or more times (144 utterances). From these utterances, we removed items which were unintelligible, contained laughter, background noise and truncated words. This yielded 72 candidates. Finally, we selected stimuli in which the *well* was followed by a pause leaving a set of 18 utterances. For the experimental stimuli, we chose two carrier phrases from the set which signals agreement or disagreement lexically: *Well no* (negative polarity) and *Well yes* (positive polarity). We additionally used *Well I don't know* as a filler item to represent an utterance lexically signalling neither agreement nor disagreement (neutral polarity). We used the remaining 18 utterances synthesised with the features from the NXT speakers as additional filler items in the experiment.

As mentioned, various prosodic features, such as duration and F_0 can be controlled in FastPitch. To create stimuli, we extracted the prosodic features (LC0-LC2 and duration) of our carrier utterances and used these features to synthesise each text. To create prosodic variation, we used the prosodic features from the closest sample to the cluster centroid in Section 2.1.4. We adapted the duration of the silence between each *well* and its following context to be approximately 13 frames long. The

[3] https://github.com/NVIDIA/NeMo

experimental stimuli therefore consisted of 20 positive-polarity, 20 negative-polarity utterances which were generated using 20 *wells*, one from each cluster centroids. The fillers consisted of 20 neutral polarity utterances generated using 20 *wells* and 18 utterances from NXT with differing textual content.

3.3. Participants and Task

Participants were recruited through Prolific[4] and were selected to reside in the US, have English as a first language with no hearing impairments. We recruited 46 participants in total. Participants were presented with 78 stimuli, in a random order. On each trial, the audio of one stimulus and a rating scale were presented. As each stimulus was presented, participants were asked *What degree of agreement/disagreement is the speaker expressing?* The instructions stated that a rating of 1 represents *complete disagreement*, 99 represents *complete agreement* and 50 represents *neither agreement nor disagreement*.

3.4. Statistical Testing

We treated the positive-polarity and negative-polarity utterances as separate datasets for statistical testing. We used linear mixed-effects models using the lme4 package [34] in R to test the effect of the four prosodic features from the cluster centroids on the rating of agreement or disagreement. We use the maximum random effects structure of the design which consists of a random-effect for participant. Our dependent variable is the rating of agreement between 1-99 which we treat as continuous. Our fixed effects consist of the acoustic features which were used to cluster the *well* dataset (see section 2.1.4), however because the coefficients are used to *condition* the model, and will show differences in output depending on neighbouring words, we extracted the acoustic features again from the synthesised stimuli.

```
agreement-rating ~ Legendre 0 + Legendre
1 + Legendre 2 + duration + (1|listener)
```

4. Results

Two participants were removed, one for having issues with playing the audio and one for indicating in the post-question that they only focused on the lexical content of the stimuli. The results for the positive-polarity carrier utterances are found in Table 3. The mean rating for these stimuli was 75.50 (SD=14.58). We found that duration was a significant predictor of agreement and had a negative relationship on the agreement rating. As duration increases by one unit, the level of agreement expressed decreases. This might signal hesitation to the listener and seems to decrease the positive polarity of *yes*. We observe that when F_0 height (LC0) increases by one unit, the agreement increases by 0.62 and when the LC2 increases by one unit, the agreement rating also increases by 0.61. This suggests that these values positively strengthen the assertion.

The results for the negative-polarity carrier utterances are found in Table 3. The mean rating for the negative-polarity stimuli was 18.03 (SD=14.71). Similar to the results for positive-polarity, duration is a significant predictor of agreement, but has a *positive* relationship on agreement. As duration increases by one unit, the level of agreement expressed increases. This again suggests that an increase in duration shifts the polarity of the utterance. Again, we observe that both the LC0 and LC2 exhibit a positive relationship on the level of agreement. Altogether this suggests that these features weaken the negative polarity of this carrier utterance. It should be noted that during statistical analysis, we observed heavy tails in the residual distributions. Though these models have been found to be robust to violations of the normality of residuals, the results warrant careful interpretation [35].

[4] https://www.prolific.com

Table 3: *Experimental Results per Polarity Type*

	Fixed Effect	β	St. Err	CI	p-value
Positive	LC0	0.62	0.25	0.13 - 1.11	< 0.05
	LC1	-0.04	0.14	-0.31 - 0.24	> 0.05
	LC2	0.61	0.24	0.14 - 1.09	< 0.05
	Duration	-7.13	2.79	-12.59 - 1.67	< 0.05
Negative	LC0	0.75	0.35	0.06 - 1.44	< 0.05
	LC1	-0.09	0.17	-0.43 - 0.24	> 0.05
	LC2	0.72	0.35	0.04 - 1.40	< 0.05
	Duration	14.35	2.79	8.89 - 19.81	< 0.01

5. Discussion and Conclusions

In this study, we presented a data-driven method for exploring prosodic patterns on the discourse marker *well* by clustering information about the duration and F_0 contour of unlabelled data. We used the centroids of the clusters to create synthetic stimuli for a listening test which tested the effect of the prosodic realisation of *well* on the degree of agreement expressed by the speaker. We found a consistent pattern regarding the effect of duration on both positive and negative short utterances. When *well* is realised with a longer duration, the rating tends away from disagreement for negative utterances, and away from agreement for positive utterances. This suggests a form of hedging for negative utterances and reluctant agreement (p 207 [36]) in the case of the positive utterance. Future work will aim to expand this study, by including more versions of the carrier utterances, to investigate whether this effect is consistent across different realisations of the context. Further, we used the prosodic features of the sample closest to the centroid of each cluster, but future work will sample more realisations from each cluster, to generate more variation and to validate our clustering.

We presented an example perceptual study to demonstrate how we can start to uncover pragmatic patterns in messy data, but this method could be applied to any other discourse marker. This method can be used by linguists, who want to create stimuli for experiments to test pragmatic theories, and by speech synthesis practitioners, who would like to explore the variation in their data. Using found data and automatic tools to filter and label the data does not have perfect accuracy, but we hope nonetheless that such corpora can be used by speech technologists to do data exploration in the hope of finding new insights, and to generate new hypotheses. Exploring the data in this way is not meant to be strictly confirmatory of linguistic theory, and as such this work was largely exploratory in nature, and theory neutral. As dialogue systems move towards expressivity at a faster pace than our knowledge about pragmatics and conversational speech phenomena, it is important that we develop methods which can uncover these important patterns. In conclusion, we propose a methodology that enables discerning knowledge and hypotheses about pragmatic functions from large amounts of found data, and validate them with perceptual experiments.

6. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 859588, the Swedish Research Council project Perception of speaker stance (VR-2020-02396), and the Riksbankens Jubileumsfond project CAPTivating (P20-0298).

7. References

[1] É. Székely, G. E. Henter, and J. Gustafson, "Casting to Corpus: Segmenting and Selecting Spontaneous Dialogue for TTS with a CNN-LSTM Speaker-dependent Breath Detector," in *Proc. ICASSP*, 2019, pp. 6925–6929.

[2] L.-W. Chen, S. Watanabe, and A. Rudnicky, "A vector quantized approach for text to speech synthesis on real-world spontaneous speech," in *Proc. AAAI'23*, vol. 37, 2023, pp. 12 644–12 652.

[3] T. Saeki, G. Wang, N. Morioka, I. Elias, K. Kastner, A. Rosenberg, B. Ramabhadran, H. Zen, F. Beaufays, and H. Shemtov, "Extending Multilingual Speech Synthesis to 100+ Languages without Transcribed Data," in *Proc ICASSP 2024*. IEEE, 2024.

[4] C. Figueroa, A. Adigwe, M. Ochs, and G. Skantze, "Annotation of Communicative Functions of Short Feedback Tokens in Switchboard," in *Proc. LREC*, Marseille, France, 2022.

[5] É. Székely, G. Eje Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," in *Proc. SSW 10*. ISCA, 2019, pp. 245–250.

[6] R. Dall, M. Tomalin, M. Wester, W. Byrne, and S. King, "Investigating automatic & human filled pause insertion for speech synthesis," in *Proc. Interspeech 2014*. ISCA, Sep. 2014, pp. 51–55.

[7] M. Wester, M. Aylett, M. Tomalin, and R. Dall, "Artificial personality and disfluency," in *Proc. Interspeech 2015*. ISCA, 2015, pp. 3365–3369.

[8] S. Andersson, J. Yamagishi, and R. A. J. Clark, "Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis," *Speech Communication*, vol. 54, no. 2, pp. 175–188, 2012.

[9] Z. Hodari, O. Watts, and S. King, "Using generative modelling to produce varied intonation for speech synthesis," in *Proc. SSW 10*, 2019, pp. 239–244.

[10] J. Hirschberg and D. Litman, "Empirical Studies on the Disambiguation of Cue Phrases," *Computational Linguistics*, vol. 19, no. 3, 1993.

[11] J. Heritage, "Well-prefaced turns in English conversation: A conversation analytic perspective," *Journal of Pragmatics*, vol. 88, pp. 88–104, Oct. 2015.

[12] A. H. Jucker, "The discourse marker well: A relevance-theoretical account," *Journal of Pragmatics*, vol. 19, no. 5, pp. 435–452, May 1993.

[13] T. I. Sakita, "Stance management in oral narrative: The role of discourse marker *well* and resonance," *Functions of Language*, vol. 24, no. 1, pp. 65–93, 2017.

[14] M. E. Helt, "Discourse marker and stance adverbial variation in spoken American English: A corpus-based analysis," Ph.D., Northern Arizona University, United States – Arizona, 1997.

[15] A. Popescu-Belis and S. Zufferey, "Automatic identification of discourse markers in dialogues: An in-depth study of like and well," *Computer Speech & Language*, vol. 25, no. 3, pp. 499–518, 2011.

[16] V. E. Michilsen, "On the relation between the prosody and discourse functions of well," BA Thesis, Utrecht University, 2019.

[17] A. Reece, G. Cooney, P. Bull, C. Chung, B. Dawson, C. Fitzpatrick, T. Glazer, D. Knox, A. Liebscher, and S. Marin, "The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation," *Science Advances*, vol. 9, no. 13, 2023.

[18] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[19] M. Lavechin, M. Métais, H. Titeux, A. Boissonnet, J. Copet, M. Rivière, E. Bergelson, A. Cristia, E. Dupoux, and H. Bredin, "Brouhaha: Multi-Task Training for Voice Activity Detection, Speech-to-Noise Ratio, and C50 Room Acoustics Estimation," in *Proc. ASRU*. Taipei, Taiwan: IEEE, 2023, pp. 1–7.

[20] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*. ISCA, 2017, pp. 498–502.

[21] R. Dall, S. Brognaux, K. Richmond, C. Valentini-Botinhao, G. E. Henter, J. Hirschberg, J. Yamagishi, and S. King, "Testing the consistency assumption: Pronunciation variant forced alignment in read and spontaneous speech synthesis," in *Proc. ICASSP*. Shanghai: IEEE, 2016, pp. 5155–5159.

[22] Y. Jadoul, B. Thompson, and B. d. Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[23] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 6.1.38)," Jan. 2021. [Online]. Available: http://www.praat.org

[24] C. de Looze and S. Rauzy, "Automatic Detection and Prediction of Topic Changes Through Automatic Detection of Register variations and Pause Duration," in *Proc. Interspeech 2009*, 2009.

[25] E. Grabe, G. Kochanski, and J. Coleman, "Connecting Intonation Labels to Mathematical Descriptions of Fundamental Frequency," *Language and Speech*, vol. 50, no. 3, pp. 281–310, Sep. 2007.

[26] M. Zellers and R. Ogden, "Exploring Interactional Features with Prosodic Patterns," *Language and Speech*, vol. 57, no. 3, pp. 285–309, 2014.

[27] U. D. Reichel, "Data-driven Extraction of Intonation Contour Classes," in *Proc. 6th ISCA Workshop on Speech Synthesis,*, 2007.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] A. Łańcucki, "FastPitch: Parallel Text-to-speech with Pitch Prediction," Feb. 2021, arXiv:2006.06873 [cs, eess].

[30] J. O'Mahony, N. Corkey, C. Lai, E. Klabbers, and S. King, "Hierarchical intonation modelling for speech synthesis using legendre polynomial coefficients," in *Proc. Speech Prosody 2024*, 2024, p. (to appear).

[31] I. Keith and J. Linda, "The LJ Speech Dataset," 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[32] D. Kontogiorgos, V. Avramova, S. Alexanderson, P. Jonell, C. Oertel, J. Beskow, G. Skantze, and J. Gustafson, "A Multimodal Corpus for Mutual Gaze and Joint Attention in Multiparty Situated Interaction," in *Proc. LREC 2018*, 2018, pp. 119–127.

[33] 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, no. 4, pp. 387–419, 2010.

[34] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, pp. 1–48, 2015.

[35] H. Schielzeth, N. Dingemanse, S. Nakagawa, D. Westneat, H. Allegue, C. Teplitsky, D. Réale, N. Dochtermann, L. Z. Garamszegi, and Y. Araya-Ajoy, "Robustness of linear mixed-effects models to violations of distributional assumptions," *Methods in Ecology and Evolution*, vol. 11, no. 9, pp. 1141–1152, 2020.

[36] K. Aijmer, "Well in an English-Swedish and English-French Contrastive Perspective," in *Researching Sociopragmatic Variability: Perspectives from Variational, Interlanguage and Contrastive Pragmatics*, K. Beeching and H. Woodfield, Eds. London: Palgrave Macmillan UK, 2015, pp. 201–229.