Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

# Hierarchical Intonation Modelling for Speech Synthesis using Legendre Polynomial Coefficients

*Johannah O'Mahony*[1], *Niamh Corkey*[1], *Catherine Lai*[1], *Esther Klabbers*[2], *Simon King*[1],

[1]Centre for Speech Technology Research, United Kingdom
[2]ReadSpeaker, The Netherlands

johannah.o'mahony@ed.ac.uk

## Abstract

Synthetic speech quality is now close to parity with human speech for isolated read speech utterances. There has therefore been a resurgence of interest in using speech synthesis for speech science research. However, many speech synthesis models lack control over prosody. The few models that are controllable do not use interpretable control values or controls that relate to prosodic theory. We present a model that enables control, by conditioning on a hierarchical Legendre polynomial representation of $F_0$ at the phrase and word levels. The polynomial coefficients are data-driven but linguistically-motivated and have been used in previous studies of pitch accents and phrase contours. The coefficients are interpretable in their characterisation of the $F_0$ contour because they describe mean $F_0$, slope, and convexity. We demonstrate sufficient control of $F_0$ to produce speech that is intonationally similar to a reference sample. Objective and subjective evaluations are used to compare our Legendre-conditioned model to a baseline, to a model conditioned on categorical prosodic features, and to an oracle model conditioned on ground-truth $F_0$. Our model has lower $F_0$ prediction error and higher correlation with ground-truth. Future work aims to apply these features to conversational speech, by learning polynomial coefficients from large speech corpora.
**Index Terms**: speech synthesis, intonation modelling, prosody control, interpretability

## 1. Introduction

Due to the increase in Text-to-Speech (TTS) quality, there have been renewed calls for the use of TTS in speech science [1], for example in prosodic research. One important requirement for using TTS is the ability to *control* specific acoustic features [1]. In this work, we focus on the controllability of $F_0$. Many current TTS models have the ability to control $F_0$, on the phone- [2] or frame-level [3]. However, controlling $F_0$ phone-by-phone is not as useful as control via more abstract, interpretable, or theoretically-relevant representations of $F_0$ used in speech science, which is desirable [4]. Further, controlling $F_0$ phone-by-phone does not account for the hierarchical structure of intonation, for example distinguishing word-level accentual features from phrase-level features such as declination. It is desirable to have a model that can be controlled hierarchically, with features describing the *shape* of the $F_0$ contour, and which can also be linked to intonation theory.

One representation of $F_0$ contour shape, which has been used in previous linguistic research to validate prosodic annotations [5, 4], is to fit the Legendre series of orthogonal polynomials, up to a certain order, to an $F_0$ contour. Each polynomial is multiplied by a specific *coefficient* before the resulting polynomials are summed [4]. After fitting these polynomials to $F_0$, we can interpret the coefficients. The first three represent $F_0$ height, slope, and convexity [5]. In this paper, we model the $F_0$ contour hierarchically using a linear regression (i.e. slope) on each phrase, and the first three coefficients of a third-order Legendre polynomial on each prosodically-prominent word. Specifically, we:

1. investigate whether conditioning a FastPitch [2] TTS model using phrase-level slope coefficients and word-level Legendre polynomial coefficients on prominent words is sufficient to provide controllability of the $F_0$ contour, as measured by similarity to a reference recording;
2. compare the method above to an alternative which uses categorical prominence and boundary markers [6];
3. compare results to the baseline FastPitch model and to a model conditioned on ground-truth mean $F_0$ per phone.

## 2. Previous Work

To use TTS in prosodic research, it is desirable to have controllable synthesis using a specification of $F_0$ which is of use to speech technologists, but can also be linked to prosodic theory [4]. Recent methods of controllability, such as Global Style Tokens, involve unsupervised representation learning, e.g., [7, 8]. However the resulting representations are not linked to linguistic theory (and in some cases, not even guaranteed to *only* represent prosody) and therefore meaningful control of such models for prosodic research is difficult.

Other methods of prosody control have focused on categorical representations of prominence and phrase boundaries. For example, [6] used the Continuous Wavelet Transform (CWT) to combine $F_0$, duration, and intensity in order to label prominence and boundaries on each word in an utterance. By conditioning a TTS model on such labels, the placement of prominence and boundaries could be controlled. Such methods, however, do not offer control of fine-grained aspects of $F_0$ realisation, such as pitch accent *shape* or boundary tone realisation.

Previous work on intonation contour modelling includes sequential models, such as the PaIntE model [9], which models the $F_0$ contour shape using two sigmoid functions. This has been applied to both TTS intonation modelling and linguistic research [10]. Hierarchical models, modelling phrase and accent components include [11] and [12]. CoPaSul [11], developed for modelling and linguistically-analysing the $F_0$ contour, combines a linear phrase component with word-level polynomial stylisation. It has not been applied to TTS modelling. Here, we take a similar approach to CoPaSul, but we do apply our method directly to TTS. We use a linear phrase component and a word-level Legendre polynomial component fitted to prominent words. Legendre Polynomials are able to characterise word-level prominence [5], and have been used in


10.21437/SpeechProsody.2024-208

studies of prosody at the word- and phrase- or utterance-level [13, 14, 15, 16] and for evaluating voice mimicry automatically [17]. Legendre Polynomials have been used previously in non-neural TTS [18, 19, 20]. More recently, Legendre polynomial coefficients were used to control the phrase realisation of an utterance in FastPitch [21]. Modelling $F_0$ on just one level, such as the phrase, will only give a coarse specification of an utterance's $F_0$ contour, and fails to capture the hierarchical relationship between phrases and their constituent words. In our work, we therefore use both phrase-level and word-level components.

# 3. Method

The goal of the current work is to validate that Legendre Polynomials provide an adequate representation of the $F_0$ contour. Our method involves performing TTS, but with prosody *transferred* from a reference recording rather predicted from the input phone sequence. Intonation is transferred using various representations of $F_0$, including Legendre Polynomials. Duration is directly copied, per phone, from the reference.

## 3.1. Data

We use the LJ Speech corpus[1], comprising 13 100 utterances read by a female US English speaker, to train all models.

### *3.1.1. Pitch accent and boundary detection*

We first aligned the data with the Montreal Forced Aligner [22] to obtain word and phone alignments. To identify prominent words and phrase boundaries, we used the Wavelet Prosody Toolkit[2] [23], which uses duration, $F_0$, and intensity. The Continuous Wavelet Transform (CWT) is calculated on the combined signals, resulting in a value for the strengths of prominences and boundaries. To identify boundaries, we used the sum of the $F_0$, intensity, and duration signals with weights 1.0, 1.0, and 0.5 respectively. For prominence, we used the product of the signals with weights of 1.0, 0.5, and 1.0 respectively. Discretising the resulting values at a particular threshold results in a categorical prominence or boundary label. Carefully articulated speech, such as read news, has between 52% and 54% of words accented [24]. We therefore used a value of 53% as a heuristic, and discretised our prominence at the 47th percentile, computed per-utterance. We chose a boundary threshold of 1.0 as this was the minimum value over all utterance-final words.

### *3.1.2. Pitch Extraction*

We used the Praat [25] autocorrelation pitch ($F_0$) estimation algorithm to obtain the $F_0$ curve, via the Parselmouth python package [26]. We first estimated $F_0$ with the default parameters: a floor of 75 Hz and a ceiling of 600 Hz. We then calculated the optimum floor and ceiling using the method in [27] and re-estimated the pitch contour. $F_0$ was transformed to semitones relative to the speaker's $F_0$ median across the entire corpus. We deleted $F_0$ values more than 2.5 standard deviations away from their utterance's mean.

### *3.1.3. Slope and Legendre Polynomial Coefficients*

To extract hierarchical phrase- and word-level features, we use the phrase boundaries and prominent words found in Section 3.1.1. Our phrase-level feature is the $F_0$ slope, found using linear regression. Each phrase receives a single slope value. For our word-level Legendre Polynomial features, we first identify the prominent words in the utterance. We then remove the effect of the previously-estimated slope by detrending the $F_0$ values: subtracting the value of the fitted regression line from all individual $F_0$ values in that phrase. We then normalise the times stamps in the words between $[-1, 1]$, and fit a third-order Legendre polynomial to them. Our word-level features comprise the *first* three coefficients of that polynomial, and represent height, slope, and convexity respectively. Non-prominent words and silences receive zero values. An example utterance with fitted slopes and word-level Legendre polynomials can be seen in Figure 1.

## 3.2. Models

All our models are variants of the FastPitch TTS model [2] using our fork of the original code[3]. FastPitch is a non-autoregressive transformer-based model that takes a phone sequence as input and predicts a mel-spectrogram. The model contains *variance adapters*, trained to explicitly predict $F_0$ and energy of each input phone. The model also explicitly predicts the duration of each input phone. When training the model, ground-truth values for $F_0$ and energy are used both to condition the model's decoder and to train the variance adapters. During inference, either the model's predictions *or* user-provided values can be used, the latter providing a means of control.

### *3.2.1. Legendre Coefficient Conditioning*

To use externally-provided phrase-level slope and word-level Legendre polynomial coefficients to control $F_0$, we modified the FastPitch architecture: Phrase-level slope values are upsampled to the number of input phones in the corresponding phrase within an utterance. This results in a tensor $S$ with shape [phone sequence length, 1]. The word-level features consist of three Legendre coefficients for each word (the coefficients are set to 0 for non-prominent words). These are upsampled to the number of input phones in the corresponding word, for each word in the utterance, resulting in a tensor $L$ with shape [phone sequence length, 3]. $S$ and $L$ are each passed through a linear layer to project them to the shape of the encoder output [phone sequence length, 384] then summed to the encoder output before being passed to the variance adapters.

### *3.2.2. Categorical Prominence Conditioning*

For our categorically-conditioned model, we conditioned a variant of FastPitch model on labels extracted as described in section 3.1.1. This model receives two externally-provided features. $P$ is a sequence of prominence labels, one for each word in the utterance (1: non-prominent, 2: prominent, 3: silence, 0: padding). $B$ is a sequence of phrase boundary labels, one for each word (1: phrase-internal, 2: prosodic phrase-final, 3: silences, 0: padding). $P$ and $B$ are each upsampled to the number of phones in the corresponding words, resulting in tensors of shape [phone sequence length, 1]. For each feature an embedding table was instantiated with an embedding dimension matching the size of the encoder output (384). Both $P$ and $B$ were passed to their respective embedding tables. The resulting embeddings were summed to the encoder output before being passed to the variance adapters.

[1] https://keithito.com/LJ-Speech-Dataset

[2] https://github.com/asuni/wavelet_prosody_toolkit/tree/master

[3] https://github.com/johannahom/FastPitches

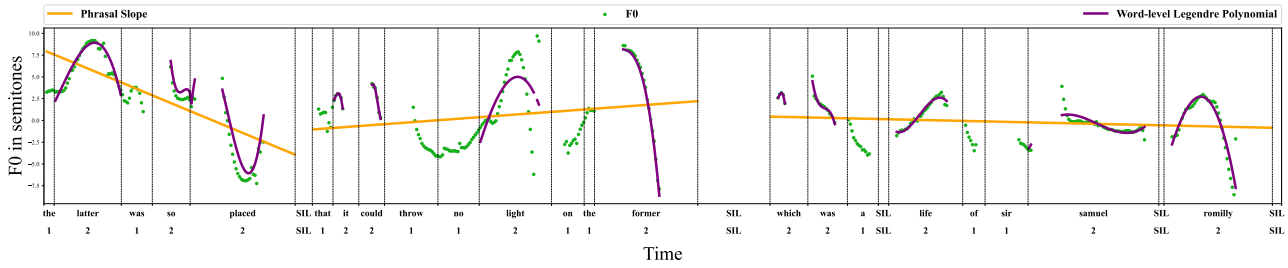


Figure 1: *$F_0$ with fitted slope and Legendre Polynomials LJ013-0179. The x-axis shows both word alignments and prominence category.*

### 3.3. Training

For the experiments described below, we trained four models.
**BASELINE** is an unmodified FastPitch model, which receives no additional conditioning inputs. It predicts $F_0$ using the usual variance adapter. As for all other models, ground-truth phone durations are used rather than predicted ones.
**ORACLE-GT-F0** is a gold standard, identical to **BASELINE** except that it uses an externally-provided ground-truth $F_0$ value per phone (rather than the value predicted by the variance adapter). We expect the gold standard will beat **LEGENDRE**, but hypothesise that the difference in listener preference will be small, thus validating that our much sparser and interpretable representation of $F_0$ is adequate.
**LEGENDRE** is the model from Section 3.2.1 conditioned on a slope parameter for each prosodic phrase and a set of three Legendre polynomial coefficients for each prominent word.
**LEGENDRE** might learn which words are prominent simply by virtue of them receiving non-zero values for polynomial coefficients vs. zero values on non-prominent words, and learn where prosodic phrase locations from the presence/absence of slope values. To confirm that **LEGENDRE** is actually using slope and coefficient *values* to predict better $F_0$, we introduce the **CATEGORICAL** model.
**CATEGORICAL** is the model from Section 3.2.2 conditioned on CWT-derived prominence and boundary labels. Prominent words are those same words that receive Legendre polynomials in **LEGENDRE**. Prosodic phrase boundaries are in the same places that the **LEGENDRE** model receives a new slope value.

For each of the models, the same 12691 utterances were used for training[4] with 129 for the measuring validation loss. Chapter 50 was reserved as test material, comprising 278 utterances. Each model was trained on a single GPU for 500 epochs.

## 4. Evaluation

Our evaluation method involves prosody transfer from an original naturally-spoken reference recording of the same text as that being synthesised. Because we are only interested in $F_0$, all models perform synthesis using ground-truth phone durations from the reference. Each model (except **BASELINE**) receives externally-provided conditioning, derived from the intonation of the reference recording: **ORACLE-GT-F0** receives the reference recording's exact $F_0$ value, per phone; **LEGENDRE** receives slopes and polynomial coefficients, fitted to the reference recording's $F_0$ contour; **CATEGORICAL** receives prominence and boundary labels, derived from the reference recording via the CWT.

[4]LJ019-0167 and LJ011-0050 removed due to processing issues

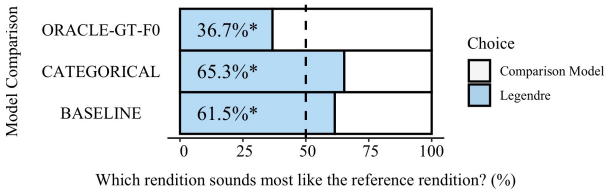


Figure 2: *Listeners' pairwise preferences between models*

For both objective and subjective evaluation, 50 utterances[5] were randomly selected from the test material. All generated mel-spectrograms were vocoded using the same pre-trained NeMo Hifi-GAN vocoder[6]

### 4.1. Subjective Evaluation

We conducted three separate listening experiments. The task in each experiment was to listen to the original natural (ground-truth) recording of a test sentence, and make a forced-choice between renditions from a pair of models. We asked listeners *"Which rendition sounds most like the reference rendition?* The three experiments compared the following pairings of models: **LEGENDRE** vs. **BASELINE**; **LEGENDRE** vs. **ORACLE-GT-F0**; **LEGENDRE** vs. **CATEGORICAL**. Within each experiment, every listener was presented with the same 50 pairs. The presentation order of the 50 pairs, and within-pair order, was randomised per listener.

Listeners who declared English as their native language and had no known hearing impairments were recruited through Prolific[7]. After the responses were collected, we removed all listeners who did not report using headphones, had difficulty playing audio, or who finished the experiment in less time than the total audio duration, resulting in 16, 19, and 19 listeners for the three experiments respectively. For analysis, we employed binomial mixed-effects regression models with a logit-link function [28], without using predictors, which is the mixed-effects equivalence of a binomial test. We included text (because the text being synthesised will have an effect on the synthetic speech, regardless of model) and listener as random effects. Below is the formula in which *choice* denotes the stimulus that the listener selected as sounding most similar to the reference recording.

```
choice ~ 1 + (1|listener) + (1|text)
```

The results in Figure 2 show that our model **LEGENDRE** (in blue) was chosen 61.5% of the time ($\beta$=0.53 (0.63 prob),

[5]Samples `https://johannahom.github.io/SP-2024/`
[6]`https://github.com/NVIDIA/NeMo`
[7]`https://www.prolific.com`

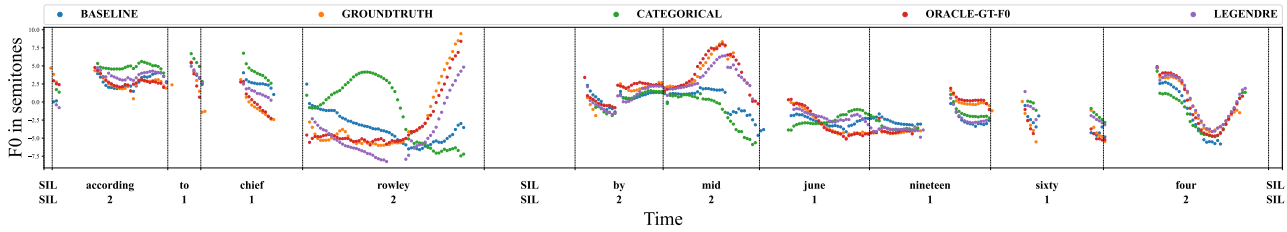


Figure 3: *$F_0$ contours of all models and ground truth for utterance LJ050-0068. $x$-axis shows both word alignments and which words received a prominence (and therefore a set of Legendre polynomial coefficients).*

CI=(0.56,0.70), $p < 0.01$) when compared to **BASELINE**. As expected, listeners preferred the gold standard **ORACLE-GT-F0** more often than our model, with **LEGENDRE** being chosen 36.7% of the time ($\beta$=-0.65 (0.34 prob), CI=(0.27,0.43), $p < 0.01$). Finally, when compared to the **CATEGORICAL** our model was chosen 65.3% of the time ($\beta$=0.68 (0.66 prob), CI=(0.61,0.72), $p < 0.01$). We can conclude that conditioning the model on phrase-level slope and word-level Legendre polynomial coefficients produces more similar-sounding intonation) than the baseline and the categorically-conditioned model. As expected, our model does not beat the gold-standard model.

Table 1: *RMSE (lower is better) and Pearson's correlation (higher is better) of polynomial coefficient and slope values between each model and ground truth, over 50 test utterances.*

| **RMSE** (no units) ↓ | | | | |
|---|---|---|---|---|
| **Model** | **Leg-0** | **Leg-1** | **Leg-2** | **Slope** |
| **ORACLE-GT-F0** | 0.72 | 1.25 | 1.50 | 0.76 |
| **BASELINE** | 1.90 | 2.79 | 2.62 | 2.40 |
| **LEGENDRE** | 1.00 | 1.58 | 1.60 | 1.31 |
| **CATEGORICAL** | 1.94 | 2.79 | 2.89 | 2.35 |
| **Pearson's correlation** ↑ | | | | |
| **Model** | **Leg-0** | **Leg-1** | **Leg-2** | **Slope** |
| **ORACLE-GT-F0** | 0.927 | 0.914 | 0.845 | 0.950 |
| **BASELINE** | 0.450 | 0.471 | 0.428 | 0.470 |
| **LEGENDRE** | 0.860 | 0.859 | 0.808 | 0.870 |
| **CATEGORICAL** | 0.477 | 0.473 | 0.327 | 0.488 |

### 4.2. Objective Evaluation

The first of two objective evaluations compares Legendre polynomial coefficients from the original recordings with those recovered from the synthetic speech. We measured Root Mean Squared Error (RMSE) and Pearson's correlation, and present results in Table 1. The second objective evaluation compares the $F_0$ values directly, again using RMSE and Pearson's correlation, with results presented in Table 2.

As hypothesised, **ORACLE-GT-F0** performs the best across all metrics. **LEGENDRE** outperforms both **BASELINE** and **CATEGORICAL** in all metrics, achieving higher correlations and lower RMSE for both polynomial coefficient values, and for $F_0$ values. This indicates that our hierarchical Legendre polynomial model offers more accurate control (in the current work, this was an intonation transfer task) than conventional pitch accent label conditioning.

The much higher correlations of polynomial coefficient and slope values for **LEGENDRE** than **BASELINE** suggest that **LEGENDRE** has indeed learned to use the provided conditioning coefficient values to predict a more accurate $F_0$ contour. Some of the difference in $F_0$ RMSE (Table 2) between **ORACLE-GT-F0** and the **LEGENDRE** model might be because **LEGENDRE** only receives non-zero coefficient values on prominent words, while **ORACLE-GT-F0** receives $F_0$ conditioning for every phone in every word.

Table 2: *RMSE and Pearson's correlation of $F_0$ between each model and ground truth, over 50 test utterances.*

| **Condition** | **RMSE** (Hz) ↓ | **Correlation** ↑ |
|---|---|---|
| **ORACLE-GT-F0** | 63.19 | 0.838 |
| **BASELINE** | 73.47 | 0.775 |
| **LEGENDRE** | 68.65 | 0.808 |
| **CATEGORICAL** | 72.35 | 0.784 |

## 5. Discussion

In this work, we have shown that using a data-driven hierarchical specification of the $F_0$ contour on the phrase-level using slope and on prominent words using Legendre polynomial coefficients outperforms traditional binary categorical conditioning: our **LEGENDRE** model produces an $F_0$ contour that sounds more similar to the reference $F_0$. As expected, our **LEGENDRE** model did not beat the **ORACLE-GT-F0** model, but our model was at a disadvantage because **ORACLE-GT-F0** received $F_0$ values for both prominent *and* non-prominent words. In future work, we plan to provide polynomial coefficients for all words, obviating the need to first detect prominent words.

Our proposed method would also allow for $F_0$ transfer from a *different* speaker. So, to further investigate whether our method provides sufficient control for use in experimental stimuli creation, we plan to transfer pitch accents from other speakers, either in found data, or from purposefully-elicited studio recordings.

Finally, to achieve TTS (i.e., no externally-provided conditioning), we plan to predict word-level Legendre coefficients from text, again comparing the results with existing accent *classification* models such as [29].

## 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.

# 7. References


[1] Z. Malisz, G. E. Henter, C. V. Botinhao, O. Watts, J. Beskow, and J. Gustafson, "Modern speech synthesis for phonetic sciences: a discussion and an evaluation," in *Proceedings of the 19th International Congress of Phonetic Sciences ICPhS 2019*, Aug. 2019, pp. 487–491.

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

[3] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *International Conference on Learning Representations*, 2021.

[4] 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.

[5] ——, "Empirical Validation of Hand-labelled Nuclear Accent Patterns," in *Proc. Speech Prosody 2006*, Dresden, Germany, 2006.

[6] A. Suni, S. Kakouros, M. Vainio, and J. Šimko, "Prosodic Prominence and Boundaries in Sequence-to-Sequence Speech Synthesis," in *Proc. Speech Prosody 2020*, 2020, pp. 940–944.

[7] X. An, Y. Wang, S. Yang, Z. Ma, and L. Xie, "Learning Hierarchical Representations for Expressive Speaking Style in End-to-End Speech Synthesis," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 184–191.

[8] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. Saurous, "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis," in *International Conference on Machine Learning*, 2018.

[9] G. Möhler and A. Conkie, "Parametric modeling of intonation using vector quantization," in *Proc. Speech Synthesis Workshop (SSW)*, 1998.

[10] A. Schweitzer, B. Möbius, G. Möhler, and G. Dogil, "The PaIntE Model of Intonation," in *Prosodic Theory and Practice*. The MIT Press, 2022, pp. 351–375.

[11] U. Reichel, "The CoPaSul intonation model," in *Sprachsignalverarbeitung, Spracherkennung und Sprachsynthese II*, Jan. 2011, pp. 341–348.

[12] M. S. Elyasi Langarani, E. Klabbers, and J. Santen, "A novel pitch decomposition method for the generalized linear alignment model," in *Proc. ICASSP*, May 2014, pp. 2584–2588.

[13] C. Lai, "Final Rises in Task-oriented and Conversational Dialogue," in *Proc. Speech Prosody 2014*. ISCA, 2014, pp. 520–524.

[14] R. Rakov, "Analyzing Prosody with Legendre Polynomial Coefficients," Ph.D. dissertation, The City University of New York, 2019.

[15] R. Rakov and A. Rosenberg, "Investigating native and non-native English classification and transfer effects using Legendre polynomial coefficient clustering," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 637–643.

[16] E. Grabe, G. Kochanski, and J. Coleman, "Quantitative modelling of intonational variation," 2004. [Online]. Available: https://ora.ox.ac.uk/objects/uuid:0730281c-dac6-4ec5-a332-22e5cd3cb8b0

[17] L. Mary, A. Babu K. K, A. Joseph, and G. M. George, "Evaluation of mimicked speech using prosodic features," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7189–7193, iSSN: 2379-190X.

[18] L. Lee, C. Tseng, and C. Hsieh, "Improved tone concatenation rules in a formant-based Chinese text-to-speech system," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 3, pp. 287–294, Jul. 1993, conference Name: IEEE Transactions on Speech and Audio Processing. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/232612

[19] C.-H. Wu, C.-C. Hsia, C.-H. Lee, and M.-C. Lin, "Hierarchical Prosody Conversion Using Regression-Based Clustering for Emotional Speech Synthesis," *Proc. IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1394–1405, 2010, conference Name: IEEE Transactions on Audio, Speech, and Language Processing.

[20] C.-C. Hsia, C.-H. Wu, and J.-Y. Wu, "Exploiting Prosody Hierarchy and Dynamic Features for Pitch Modeling and Generation in HMM-Based Speech Synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1994–2003, Nov. 2010.

[21] N. Corkey, J. O'Mahony, and S. King, "Intonation Control for Neural Text-to-Speech Synthesis with Polynomial Models of F0," in *Proc. INTERSPEECH*, Dublin, Ireland, 2023, pp. 2014–2015.

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

[23] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[24] J. Yuan, J. M. Brenier, and D. Jurafsky, "Pitch accent prediction: effects of genre and speaker," in *Proc. INTERSPEECH*. Lisbon, Portugal: ISCA, 2005, pp. 1409–1412.

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

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

[27] C. D. 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, pp. 2919–2922.

[28] 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.

[29] A. Talman, A. Suni, H. Celikkanat, S. Kakouros, J. Tiedemann, and M. Vainio, "Predicting Prosodic Prominence from Text with Pre-trained Contextualized Word Representations," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, M. Hartmann and B. Plank, Eds., Turku, Finland, 2019, pp. 281–290.