INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Intonation Control for Neural Text-to-Speech Synthesis with Polynomial Models of F0

Niamh Corkey, Johannah O'Mahony, Simon King

The Centre for Speech Technology Research, University of Edinburgh, UK
s1936986@ed.ac.uk, johannah.o'mahony@ed.ac.uk, Simon.King@ed.ac.uk

Abstract

We present a novel, user-friendly approach for controlling patterns of intonation (a fundamental aspect of prosody) within a neural TTS system. This involves concisely representing F0 contours with the coefficients of their Legendre polynomial series expansion, and implementing a model (based on FastPitch) which is conditioned on these sets of coefficients during training. At inference time the model will explicitly predict a coefficient set, or a user (eg. human-in-the-loop) can provide a target coefficient set which audibly alters the intonation of the output speech, based on just a few values. This is particularly effective for intonation transfer: where these coefficient targets are extracted from a ground truth recording, making the synthesised utterance more closely reflect the intonation of the real speaker.
Index Terms: text-to-speech, speech synthesis, intonation modelling, prosody control, prosody transfer

1. Introduction

The rise of end-to-end neural text-to-speech (TTS) models has enabled synthetic voices to sound remarkably natural and intelligible. However, most state-of-the-art TTS systems are unable to produce different prosodic renditions of a given text, thus failing to capture a key feature of natural human speech.

Approaches to combat this include prosody control/transfer, where target prosodic patterns can be manually specified or extracted from reference recordings. However, most of these techniques involve learning rich latent representations of desired prosodic features [1, 2], which is resource-intensive and can lead to source-speaker leakage.

Here, we will focus on control/transfer of one major element of prosody: intonation. We have modeled patterns of intonation simply: using the first three coefficients of the Legendre polynomial series expansion of the F0 contour for each utterance. These coefficients are interpretable, representing the average level, slope, and convexity of F0 respectively. We have adapted the FastPitch model to be conditioned on a coefficient set for each training utterance, and explicitly predict coefficients during inference. In our demo, a user can supply their own target coefficient set instead (representing desired F0 contour shape), to influence the intonation in the final synthetic speech.

2. Implementation

2.1. The Data

We trained our model on LJ Speech [3], a public domain speech dataset with a single American female speaker. The original dataset contained 13100 utterances with an average length of 6.57 s, however we found that 3-coefficient representations of these long utterances' F0 contours tend to be rather flat, and thus would not expose the model to much variety in patterns of intonation. Therefore, we rechunked this into shorter utterances based on occurences of commas and full stops in the transcriptions, to create shorter prosodic-phrase-like utterances. This resulted in 18418 utterances with an average length of 3.83 s.

We extracted F0 contours from all utterances, before taking the average value per phone and normalising values according to the speaker's mean/standard deviation of F0 (as this is consistent with FastPitch's representation of F0 contours). We then linearly interpolated the corrected contour and used the function `polynomial.legendre.legfit` from the `numpy` package to return the three coefficients of the Legendre series (of degree 2) which fits each F0 contour with least squared error. Finally, we range normalised each of the three coefficients across the dataset between -1 and 1.

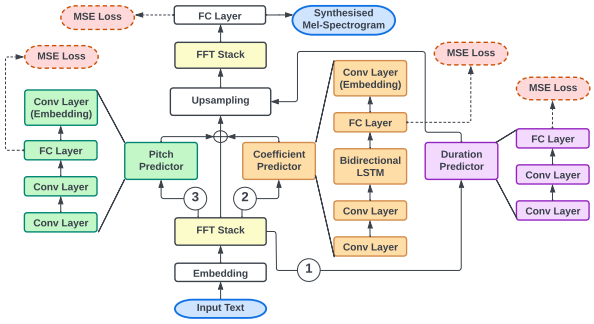

Figure 1: *Architecture of the new implementation of the FastPitch model, with 'Coefficient Predictor' added, where 'Conv' = 1D Convolutional, and 'FC' = Fully-Connected.*

2.2. The Model

We adapted FastPitch: a neural text-to-speech model which rapidly synthesises mel-spectrograms based on input text [4], and already includes explicit predictors for F0, duration and energy, allowing control over these features. The open-source Python code used as a basis for this implementation came from a fork [1] of NVIDIA's 'DeepLearningExamples' repository [5] on GitHub.

We added an explicit Coefficient Predictor to the model's architecture (as shown in Figure 1) which includes a bidirectional LSTM layer, where the prediction at the final timestep is taken as the utterance-level coefficient set. The coefficient prediction is made *before* the F0 predictions, with the hope that F0

[1] https://github.com/evdv/FastPitches

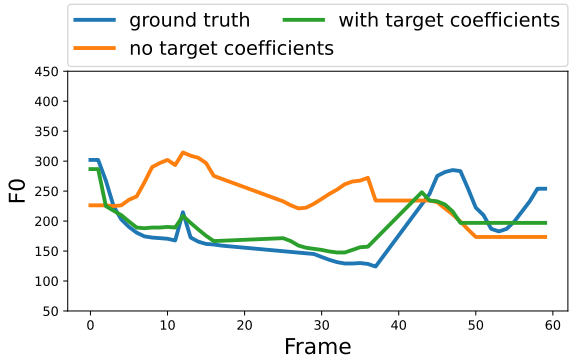

Figure 2: *F0 contours (interpolated) of ground truth recording vs synthesised versions either with ground truth coefficients provided as targets, or with no targets provided, for the text 'the little stone seal'.*

predictions can also be conditioned on the shape of intonation indicated by the coefficients. These predictions have their mean squared error loss recorded during training.

During inference, if no coefficient targets are specified, this prediction is upsampled to the maximum input sequence length in the batch, reshaped for consistency with other feature predictions, and masked by the encoder mask to reflect each different sequence length in the batch, before being embedded and summed to the current encoder output. During training, and during inference if targets are supplied, it is the ground truth/target coefficients which are upsampled, reshaped, masked, embedded and summed to the encoder output.

The final model was trained for 1000 epochs, with a learning rate of 0.1 and a batch size of 16, on a single GPU.

3. Intonation Transfer Results

3.1. Objective Measures

We synthesised two versions of each text in the test set (which included 736 utterances): one with coefficients from the ground truth recording supplied as targets, and one with no targets supplied. We found that average root-mean-square error between ground truth and synthesised contours was reduced when ground truth coefficient targets were used for synthesis (**47.370 Hz**) compared to when no targets were used (**55.143 Hz**). This suggests that using this method can increase the similarity in intonation between synthesised speech and human reference recordings, as exemplified in Figure 2.

3.2. Formal Listening Test

A listening test was conducted with 26 paid participants, using 50 reference recordings from the test set. Listeners were played the reference recording, followed by two synthesised versions of the same text (with-targets and no-targets, as above), and had to select which synthesised version sounded more similar to the reference in terms of pitch/intonation. The with-targets version was chosen **63.7%** of the time, which was found to be a significant preference according to a binomial mixed effects model analysis (β=0.62 (0.65 prob), CI=(0.60,0.70), $p < 0.001$). This means that this technique can audibly increase the similarity in intonation between synthesised speech and human reference recordings, proving some degree of intonation transfer ability, at least from recordings of the same speaker/text.

3.3. Informal Extension

We then tested the model's ability to transfer patterns of intonation from different speakers and different text. A new speaker was recorded producing the same text ('she said that she's your sister') as both a declarative statement (with an overall declination of F0) and a declarative question (with a characteristic rise in F0 at the end). Three coefficients were then extracted from each of the contours, creating a 'statement' target set, and a 'question' target set. The same text was then synthesised with each target set, and an informal listening test with 12 participants found that listeners were always able to correctly distinguish which version was intended to be a statement, and which a question. The same targets were then used to synthesise two versions of a different text ('it's raining in Glasgow'), and listeners were still always able to identify the statement/question correctly.

These results preliminarily suggest that this approach can be used to transfer patterns of intonation between different speakers/texts, and that the effect is obvious enough to alter the the meaning of a synthesised utterance.

4. Dicussion

We have shown that this approach offers a simple and efficient way to transfer patterns of intonation from reference recordings to synthesised speech. We hope to provide an opportunity for users to extract target coefficient sets from their own reference recordings or even hand-drawn F0 contours to allow intuitive human-in-the-loop control over synthesised intonation.

In future we hope to extend this method, for example by using the coefficients to actually reconstruct an F0 contour within the model, which should allow FastPitch's current Pitch Predictor to be fully replaced; allowing more complete control via coefficients. Anticipated future applications include using this technique to model intonation hierarchically (ie. at the word, phrase, and sentence level simultaneously), to enable more precise intonation control, and clustering coefficient sets to identify and recreate general patterns of intonation which correspond to particular speech acts/emotions.

5. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 859588.

6. References

[1] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 6945–6949.

[2] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *2018 International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[3] K. Ito and L. Johnson, "The LJ Speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[4] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 6588–6592.

[5] K. Kudrynski, "Deep learning examples: FastPitch," https://github.com/NVIDIA/DeepLearningExamples/tree/master/\PyTorch/SpeechSynthesis/FastPitch, 2020.