Joint Modeling of F0 and Duration in Deep Neural Network Based Speech Synthesis

Srikanth Ronanki, Zhizheng Wu, Robert A. J. Clark
Centre for Speech Technology Research, University of Edinburgh

Fundamental frequency(F0) and duration are two important factors in prosody, and a significant amount of work has been done to model them in statistical parametric speech synthesis. However, conventional techniques assume conditional independence between F0 and duration, and model them separately. This paper proposes an approach to jointly model the high-level behaviour of F0 contours and duration within a deep neural network framework.

1 Relation to prior work

Statistical parametric speech synthesis(SPSS) based on hidden Markov models (HMMs) [4] has flourished for many years now and offers greater flexibility to change its voice characteristics than concatenative speech synthesis approaches. Most recently, neural networks have re-emerged as a potential acoustic model for [3, 5] following their success in deep learning for ASR. However, the naturalness of synthesized speech is still generally neutral in terms of prosody and cannot compete with good unit selection systems. One likely reason for this is that pitch variation continues to be modeled locally at the frame level and these models unable to capture long-term behaviour in the pitch contour. Additionally duration is always predicted first and independently of the pitch.

The Discrete Cosine Transform (DCT) is often used to represent F0 at higher levels, which is able to compactly represent complex contours. The Continuous Wavelet Transform (CWT) has also been proposed to improve F0 within the HMM-framework [2]. Here some improvements were seen in the accuracy of F0 modeling. The work proposed in [1] explored a multi-level representation of F0 by combining both DCT and CWT transforms and modeled at different wavelet scales with each scale representing the variations in F0 contour from utterance/phrase level to phoneme level. However, all these approaches require separate models to predict state-level duration first and then use it here to predict the contour.

The novelty of this work is to jointly model duration and the F0 contour above the frame level using on deep neural networks. For this, the DCT representation of F0 contour, the duration of phone and its states are used as input features to the network. This ensures the training to learn the phoneme duration along with state duration with least deviation in between them. In the current work, we also investigate the use of bottleneck features [3] from the DNN as a richer representation of prosodic context to supplement the input. Since, the proposed approach learns the representation at phoneme level, the addition of bottleneck features provide wider context to help the network learn more accurate longer-term variations in F0 contour.

2 Proposed Joint Modeling of F0 and Duration

Let $x_i = [x_i(1), ..., x_i(d_x)]^T$ and $y_i = [y_i(1), ..., y_i(d_y)]^T$ be static input and target feature vectors of phoneme i where d_x and d_y denote the dimensions of x_i and y_i, respectively, and T denotes transposition. The input features (x_i) include binary features derived from a subset of the questions used by the decision tree clustering in the HMM system. The target features (y_i) include DCT-parameters of a phoneme F0 contour along with its duration and the time-aligned duration of its five states as shown below:

$$f_{0ERB} = \log_{10}(0.00437 * f_0 + 1) \tag{1}$$

$$c[k] = 2 * w[k] * \sum_{n=0}^{N-1} f_i[n] cos(\frac{\pi(2n+1)(k)}{2N}), \quad 0 \le k < N \tag{2}$$

where $f_i = [f_0(1), ..., f_0(t)]$, t denotes number of frames in phoneme i.

$$y_i = [c[0], ..., c[k], p_i, s_i^1, ..., s_i^5] \tag{3}$$

where $c[0], ..., c[N-1]$ represents the DCT-stylized F0 features, p_i is phoneme duration, $s_i^1, ..., s_i^5$ represents duration of states. The DNN is then trained to map the linguistic features of input text to the prosodic features of a speaker, i.e., if $D(x_i)$ denotes the DNN mapping of x_i, then the error of mapping is given by $\varepsilon = \sum \|y_i - D(x_i)\|^2$ is defined as

$$D(x_i) = \widetilde{d}(z_{n+1}) \tag{4}$$

$$z_{n+1} = d(w^{(n)} d(z_n)) \tag{5}$$

$$d(\vartheta) = a\tanh(b\vartheta), \widetilde{d}(\vartheta) = \vartheta \tag{6}$$

where n represents number of hidden layers and $w^{(n)}$ represents the weight matrix of n^{th} hidden layer of the DNN model. We tried different architectures by varying the total number of hidden layers from 3 to 6. The best architecture with minimum generation error was found out to be with 6 layers consisting of 1024 nodes in first two hidden layers and 128 nodes in the subsequent four hidden layers. Since the number of input nodes are more in number compared to output layer, the architecture [1024 1024 128 128 128 128] performed better than all other architectures during training.

In another system, a first DNN with architecture [1024 32 1024 1024 1024 1024] is used to extract 32-dimensional bottleneck features with 10 frames context and are stacked with linguistic features as input to a second DNN with architecture [6*1024] to predict prosody features. The bottleneck features represent activation at the hidden layer for each phoneme. The weights generated during DNN training are used to estimate the contour shape (c_1 - c_8) and $mean f_0$. With the help of IDCT and voiced/Unvoiced (V/UV) values from the baseline DNN, the F0 contour is reconstructed and the output F0 values are normalized to zero based on the predicted value of V/UV (if V/UV $<$ 0.5). A STRAIGHT vocoder is used to synthesize the waveform using the predicted Mel-Cepstral, BAP features from the baseline DNN's frame-by-frame mapping and F0 from the proposed method. A speech database from a British male speaker (Nick) was used in the experiments. Objective evaluation is shown in Table 1.

Table 1: Objective evaluations on Nick Database

	F0 contour(Hz)		Duration(Frames)
Method	**RMSE**	**CORR**	**RMSE**
HMM-GV	9.90	0.782	6.05
Baseline-DNN	9.34	0.812	-
DCT-DNN	9.00	0.818	5.15
DNN-DNN	8.86	0.825	5.24

We have proposed the use of joint modeling of F0 and duration in current state-of-the-art DNN-based speech synthesis. It has been shown that discrete cosine transform is a good representation of F0 contour and can be modeled using deep neural networks much better than frame level F0 modeling. Objective evaluation also suggest that these are quite effective.

3 References

[1] S. Ribeiro Manuel and R. A. J. Clark. A Multi-Level Representation of F0 Using The Continuous Wavelet Transform And The Discrete Cosine Transform. In *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2015.

[2] A. Suni, D. Aalto, T. Raitio, P. Alku, and M. Vainio. Wavelets for intonation modeling in HMM speech synthesis. In *Proc. 8th ISCA Speech Synthesis Workshop*, 2013.

[3] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[4] T Yoshimura, K Tokuda, T Masuko, T Kobayashi, and T Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *EUROSPEECH*. ISCA, 1999.

[5] H. Zen, A. Senior, and M. Schuster. Statistical Parametric Speech Synthesis Using Deep Neural Networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7962–7966, 2013.