A template-based approach for intonation generation using LSTMs

Srikanth Ronanki, Gustav Eje Henter, Zhizheng Wu, Simon King

The Centre for Speech Technology Research (CSTR), The University of Edinburgh, UK
srikanth.ronanki@ed.ac.uk

1. Abstract

The lack of convincing intonation makes current parametric speech synthesis systems sound dull and lifeless, even when trained on expressive speech data. Typically, these systems predict the fundamental frequency (F0) frame-by-frame using regression models. This approach leads to overly-smooth pitch contours and fails to construct an appropriate prosodic structure across the full utterance. In order to capture and reproduce larger-scale pitch patterns, we propose a classification-based approach to automatic F0 generation, where per-syllable pitch-contour templates (from a small, automatically-learned set) are predicted by a recurrent neural network (RNN). The use of templates mitigates the over-smoothing problem: with only six templates, we can reconstruct pitch patterns observed in the data well (small RMSE). The long memory of RNNs in principle enables the prediction of pitch-contour structure spanning the entire utterance. To construct a complete text-to-speech system, this novel F0 prediction system is used alongside separate LSTMs for predicting phone durations and remaining acoustic features. The objective results are encouraging, but listening tests with oracle reconstructions suggest that further work (beyond a simple smoothing) is necessary to reduce subjective artefacts in the template-based F0 reconstructions.

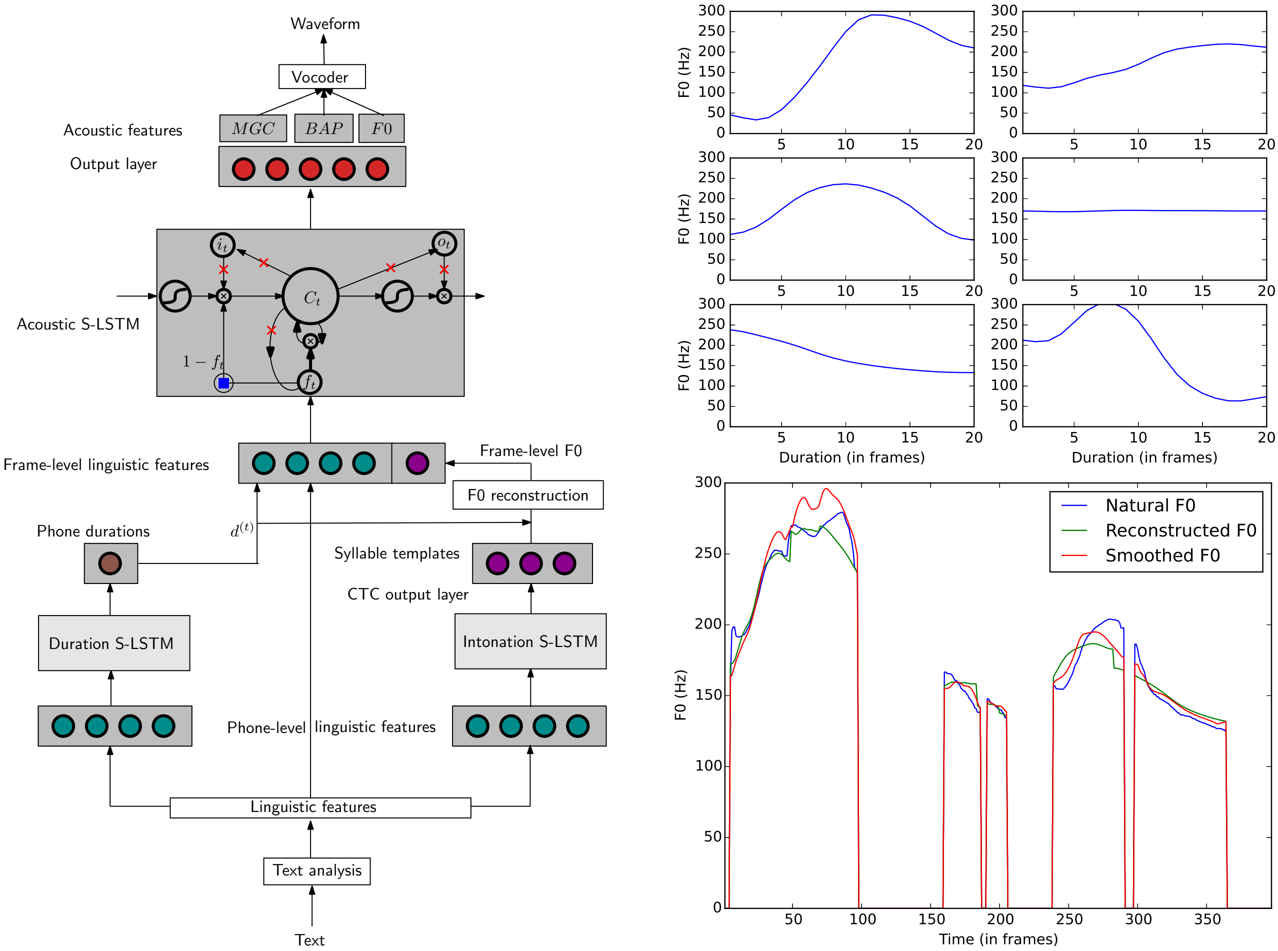

Figure 1: Schematic diagram (left) of the proposed speech synthesis system using a set of six syllable F0 templates (top right). For clarity, only a single LSTM unit is shown. The connections crossed out in red are omitted in the simplified LSTM units [1] used in this work. The bottom right plot shows raw and smoothed F0 contours reconstructed from an oracle template decomposition of natural F0.

2. References

[1] Z. Wu and S. King, "Investigating gated recurrent networks for speech synthesis," in *Proc. ICASSP*, 2016, pp. 5140–5144.