A template-based approach for speech synthesis intonation generation using LSTMs

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

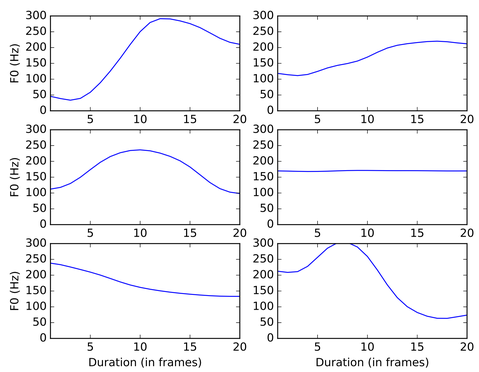
srikanth.ronanki@ed.ac.uk,simon.king@ed.ac.uk

Overview

- The lack of convincing intonation makes current parametric speech synthesis systems sound dull and lifeless
- 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.
- We propose a classification-based approach to automatic F0 generation.

Inventory of syllable F0 templates

1. Interpolate the F0 contour of each utterance and segment into syllables.
2. Apply DCT based decomposition: c_0 representing the mean over syllable, $\boldsymbol{c} = [c_1, \ldots, c_{N-1}]^\intercal$, representing the shape of the contour.
3. Perform top-to-bottom hierarchical clustering over the patterns ($\boldsymbol{c}$).



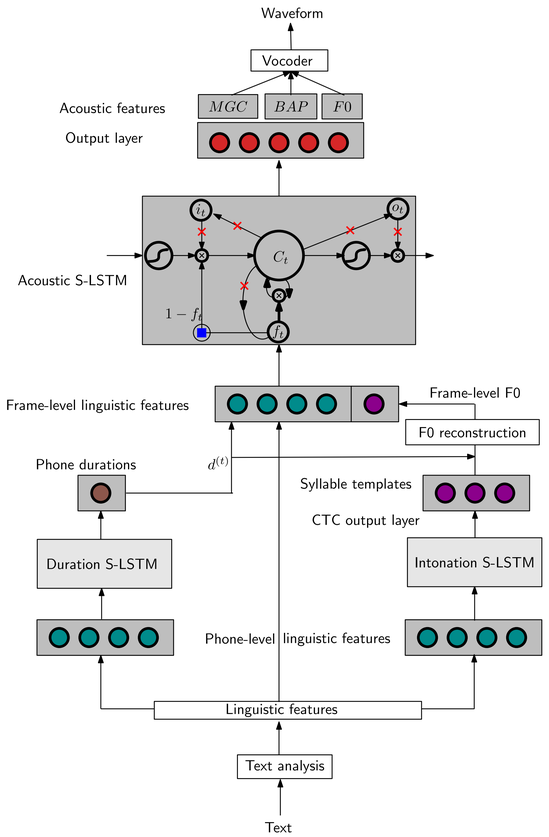
A set of six syllable F0 templates found by clustering of the data, plotted at the average F0 (180 Hz) and syllable duration (20 frames) of the speaker.

Template	1	2	3	4	5	6
Train	852	5216	1853	26725	5784	1013
Dev.	26	139	50	653	147	28
Test	65	362	106	1553	377	40
DNN	0	127	4	2247	155	0
HC	15	298	27	1676	499	18
CTC	16	200	61	1958	287	11

Template counts in the data and corresponding test-set predictions

Neural network classifiers

1. A hierarchical deep neural network classifier (HC).
   - The first DNN choses between flat and non-flat template, and then the second DNN choses among rest of the non-flat templates.
2. A simplified LSTM with a CTC output layer (CTC).
   - Connectionist temporal classification coupled with S-LSTM to predict the sequence of templates given sequence of phonemes.



Schematic diagram of the proposed speech synthesis system

Experimental Data

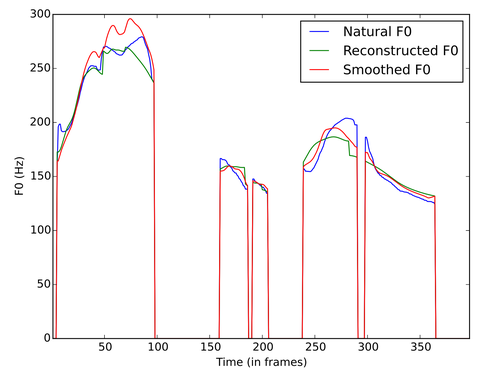
- Data: Blizzard Challenge 2016
- Language: English; Utterances: 5587; Duration: 4 Hours
- Contains 50 children's audiobooks read by a British female speaker

Results

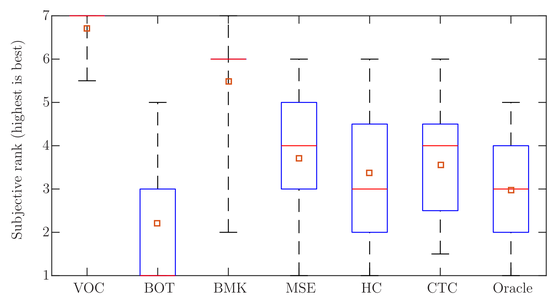
Model	Classification measures		F0 measures	
	Accuracy	F1 score	RMSE	Corr.
MSE	-	-	45.9	0.40
HC	61.1%	0.590	46.9	0.36
CTC	63.8%	0.593	46.1	0.40
Oracle	100%	1	40.8	0.58

F1 score of predicted syllable templates along with RMSE and Correlation of the predicted F0 contour

MSE: A framewise-regression baseline predicting F0 using S-LSTMS.
Oracle: Finds the matching template based on eucledian distance between syllable F0 contour and templates.



Natural F0 and reconstructed F0 contours for the sentence "Goldilocks and the three bears" at 5 ms frame rate.



Box plot of aggregate ranks from listening test

Natural Speech Technology
Edinburgh – Cambridge – Sheffield