Median-Based Generation of Synthetic Speech Durations using a Non-Parametric Approach

{ Srikanth Ronanki, Oliver Watts, Simon King, Gustav Eje Henter* }   The University of Edinburgh

Overview

- **Goal:** Modelling speech-sound durations for statistical parametric speech synthesis
  - Important for natural prosody
  - Our (Gaussian) models are wrong
- **Proposal:** Predict a *transition probability* for each frame
  - Non-parametric model (e.g., non-Gaussian distributions)
  - Median-based (not mean) generation, good for incremental speech synthesis
  - Can model acoustics+duration jointly

We use LSTM RNNs, but the idea applies to many other machine-learning paradigms and scenarios

1. Non-parametric model

Training phase builds a statistical (regression) model of durations in aligned speech+text data

- Standard approach: Phone/state-level
  - Duration and acoustic models make predictions at different intervals
  - Durations are assumed Gaussian
  - Train to min weighted mean squared error (i.e., max Gaussian likelihood)
- New approach: *Frame-level predictions*
  - Predict transition-probability $p_D(d) = P(\mathrm{Dur} = d \mid \mathrm{Dur} \geq d)$ for each frame
  - In training data, set $p_D(d) = 1$ in last frame of state/phone, 0 otherwise
  - Train to min mean squared error (MSE)
- Properties of new approach:
  - Non-parametric – can represent *any* distribution $P(\mathrm{Dur} = d)$ on positive d
  - Global MSE minimum at true $p_D(d)$
  - Predicts in parallel to acoustic model

2. Duration generation

To speak, durations are generated from model

- Standard approach: Mean-based generation
  - Mean $\widehat{d} = \mathbb{E}(\mathrm{Dur})$ can only be calculated knowing $P(\mathrm{Dur} = d)$ for *all* d
  - Hard for non-parametric distributions
- New approach: *Median-based generation*
  - Median $\widehat{d} = \min d$ s.t. $P(\mathrm{Dur} > d) < 0.5$ just requires $P(\mathrm{Dur} = d)$ for $d \leq \widehat{d}$
  - Attractive for sequential generation
  - Closer to typical (most likely) duration
  - Statistically robust to outliers
  - Can be generalised to quantile-based generation to control speech rate

Systems tested

- Phone-DNN
- Phone-LSTM
- Frame-LSTM-I
- Frame-LSTM-E

- **Phone-DNN:** Phone-level duration DNN in two-stage approach
- **Phone-LSTM:** Phone-level duration LSTM in two-stage approach
- **Frame-LSTM-*:** Frame-level duration prediction using LSTMs
- **Codebase:** Merlin [1]

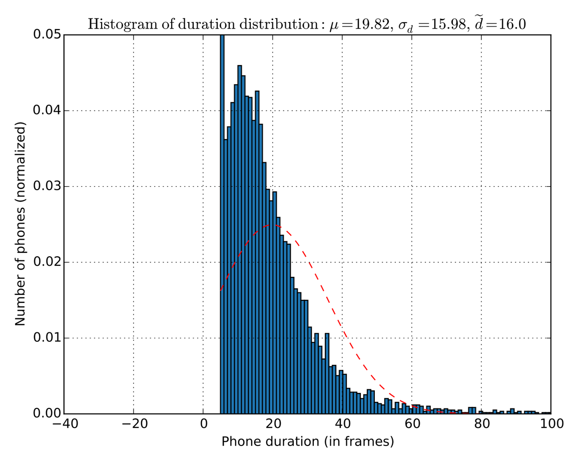

Figure 4: Duration distribution of natural durations

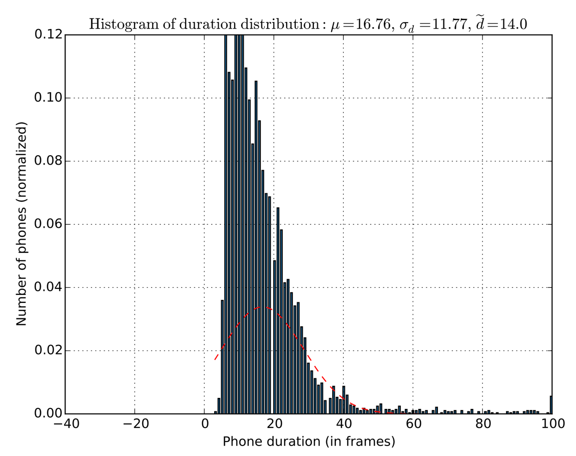

Figure 5: Duration distribution of predicted durations

Frame-level duration modeling

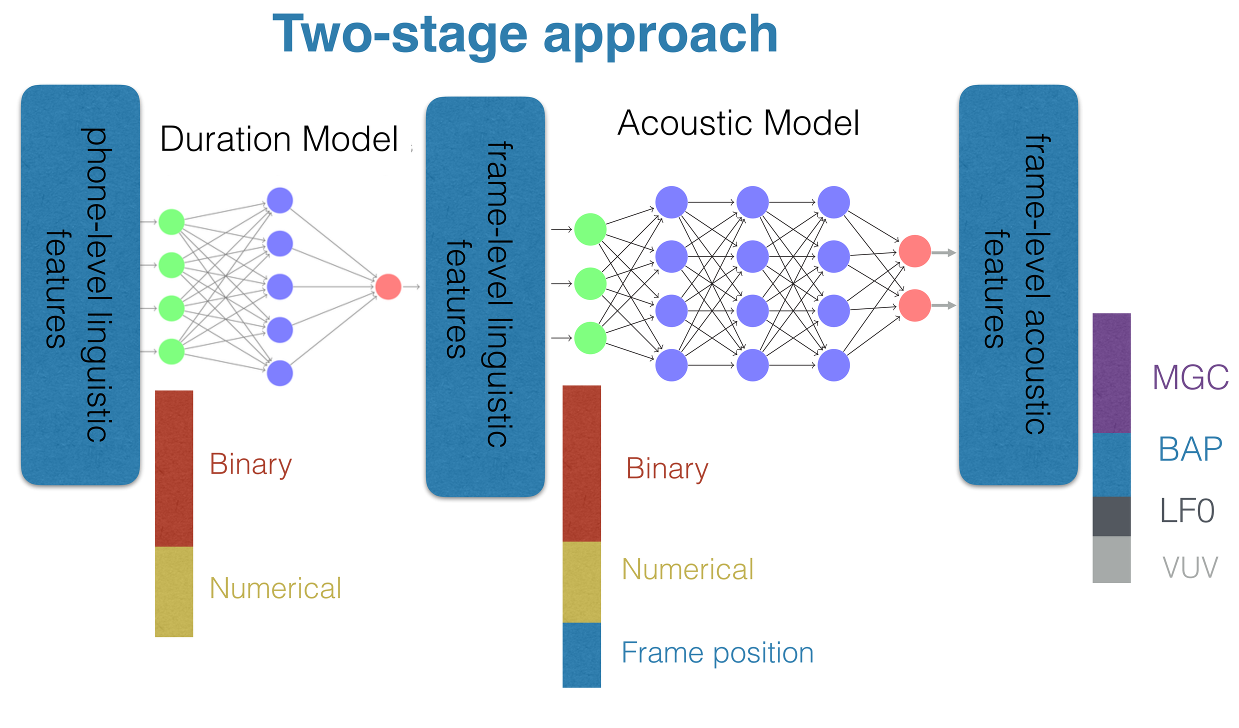

Figure 1: Schematic diagram of the two-stage approach

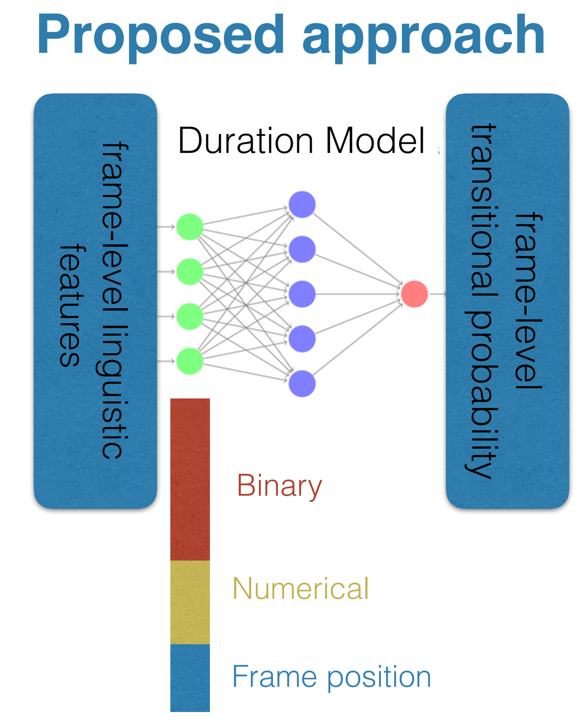

Figure 2: Schematic diagram of proposed approach

Results

- RMSE (*root mean squared error*) = Error measure minimised by true mean duration
- MAE (*mean absolute error*) = Error measure minimised by true *median* duration
- RMSE and MAE are measured in units of frames per phone
- Corr (*Pearson correlation*) = Closely related to RMSE, except higher is better

Model	RMSE	MAE	Corr.
Phone-DNN	8.037	4.759	0.750
Phone-LSTM	7.789	**4.556**	0.765
Frame-LSTM-I	8.254	4.610	0.761
Frame-LSTM-E	8.294	**4.574**	0.754

- *Phone-LSTM* always better than *Phone-DNN*
- Proposed methods closed the gap for MAE
- MAE improved for all consonant classes except plosives
- We no longer optimise for RMSE, so RMSE performance regression is expected

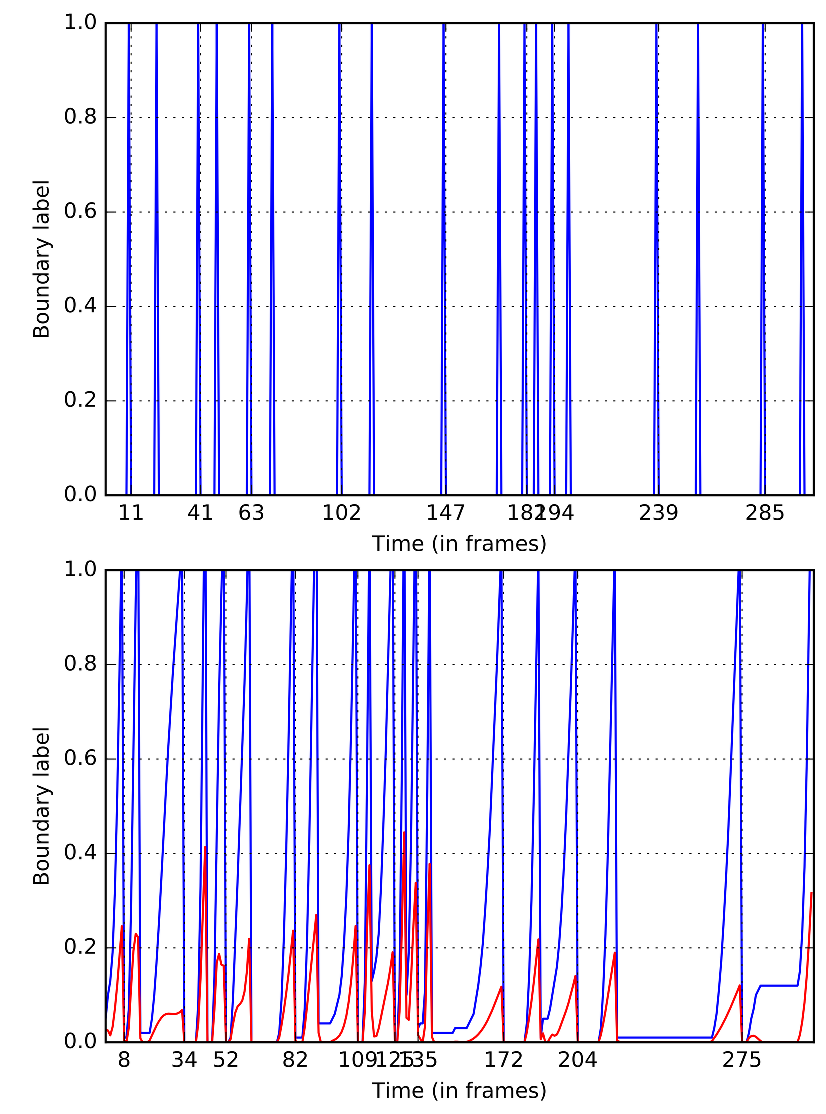

Figure 3: red: trans. prob. (π_t); *blue : scaled*(D_p)

$$(D_p = n_t \mid \boldsymbol{L}_t) = 1 - (2 * \pi_t \prod_{t'=t_0+1}^{t_0+n_t-1} (1 - \pi_{t'})) \quad (1)$$

References

[1] Zhizheng Wu, Oliver Watts, and Simon King. Merlin: An Open Source Neural Network Speech Synthesis System. In *Proc. Speech Synthesis Workshop (SSW9)*, 2016.

Future research

- Joint modelling of duration and acoustic features – a hierarchical framework to predict all features together, given phone-level linguistic features
- Conduct subjective evaluations of synthesised speech

Contact

Web http://www.srikanthr.in
Email srikanth.ronanki@ed.ac.uk
***Last author** of this paper is now at Yamagishi lab, National Institute of Informatics, Tokyo