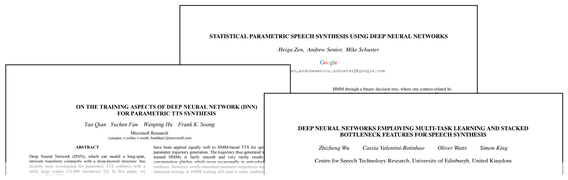
From HMMs to DNNs: where do the improvements come from?

Oliver Watts, Gustav Eje Henter, Thomas Merritt, Zhizheng Wu, Simon King ◆ *Centre for Speech Technology Research, University of Edinburgh*

Motivation

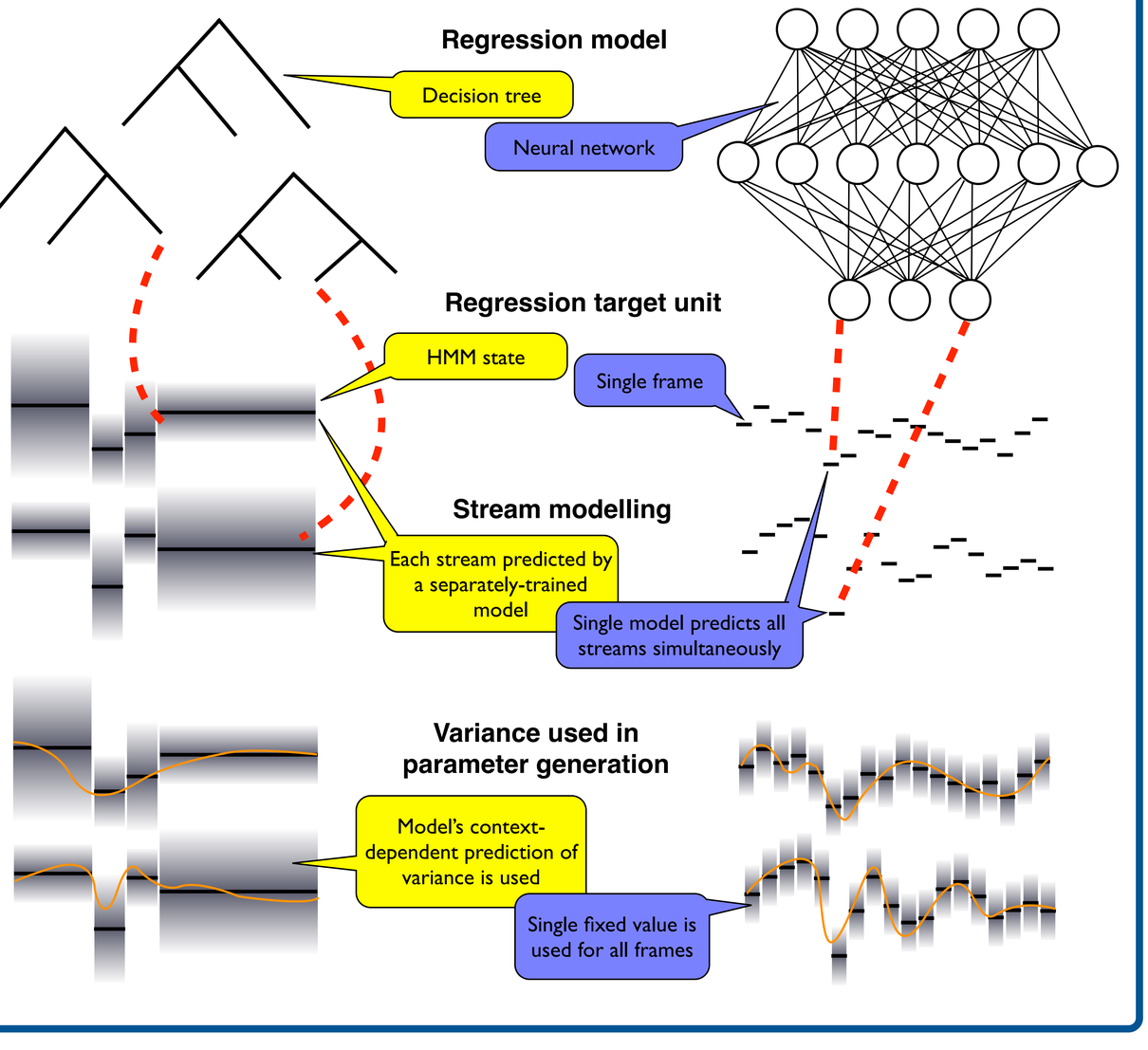
Recent revival of interest in neural networks for TTS



In this recent work:

- several configurations of DNN systems often compared
- HMM/decision-tree system generally taken as baseline
- source of improvement over baseline is hard to determine because multiple factors are simultaneously varied between systems

HTS vs. DNN-TTS



Systems built

- A range of systems was built with different combinations of the factors of interest
- Comparison of these systems allows us to attribute importance to the different factors.
- At each end of the range were standard systems
  - from the HTS demo
  - our own baseline DNN system used in previous work.
- The systems built step gradually between these endpoints
- Not all combinations were implemented (e.g. a Clustergen-type system, where decision trees operate at the frame level)

HTS public demo, with STRAIGHT

Mixture density network with single component

System	Regression model	Regression target unit	Stream modelling	Variance	Duration-derived features	Enhancement method
D1	decision tree	state	separate	context-dependent	no	GV
D2	decision tree	state	separate	context-dependent	no	postfilter
N1	neural network	state	separate	context-dependent	no	postfilter
N2	neural network	state	separate	fixed	no	postfilter
N3	neural network	state	combined	fixed	no	postfilter
N4	neural network	frame	separate	fixed	no	postfilter
N5	neural network	frame	combined	fixed	no	postfilter
N6	neural network	frame	combined	fixed	yes	postfilter

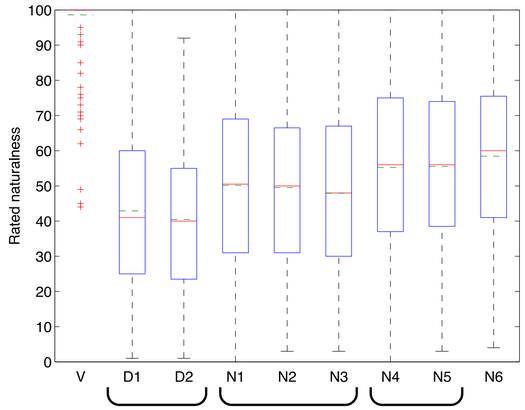
3 networks per system (mcep, log F0, band aperiodicities); 3 times as many model parameters as N3, N5, N6

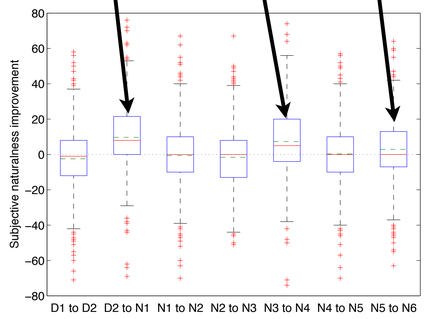
N6 uses 9 duration-derived input features; other systems use only 2

1. **Fraction through state counting forwards**
2. *Fraction through state counting backwards*
3. *Fraction through phone counting forwards*
4. *Fraction through phone counting backwards*
5. **Position of state in phone counting forwards**
6. *Position of state in phone counting backwards*
7. *Length of state in frames*
8. *Length of phone in frames*
9. *Fraction of the current phone made up by current state*

- We did not control for two factors: question set size (2926 vs. 863) and F0 modelling method (MSD vs. interpolation). Effects of these factors are currently combined with decision tree → neural network factor

Evaluation





- MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test
- 20 native English listeners
- Each listener rated two sets of 10 synthesised Harvard sentences, every set phonetically balanced

Pairwise Wilcoxon signed-rank comparisons between all systems ($\alpha = 0.05$, Holm-Bonferroni corrected) show three significant differences between groups of systems

- Two factors of approximately equal importance:
  - state → frame level modelling
  - decision tree → neural network
- Complex duration features also significantly improve naturalness (when evaluating using oracle durations)
- Enhancement method, context-dependent variance, combined vs. separate stream modelling not found to be important