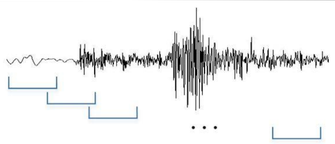
Integrated speaker-adaptive speech synthesis[1]

Moquan Wan, Gilles Degottex, Mark J.F. Gales
mw545@cam.ac.uk, gad27@cam.ac.uk, mjfg@eng.cam.ac.uk University of Cambridge

UNIVERSITY OF CAMBRIDGE
Department of Engineering

Speech Synthesis Pipeline



We also need...

/sil/ /w/ /i/ /O/ /lw/ /s/ /@U/ /n/ /i/ /d/ ...

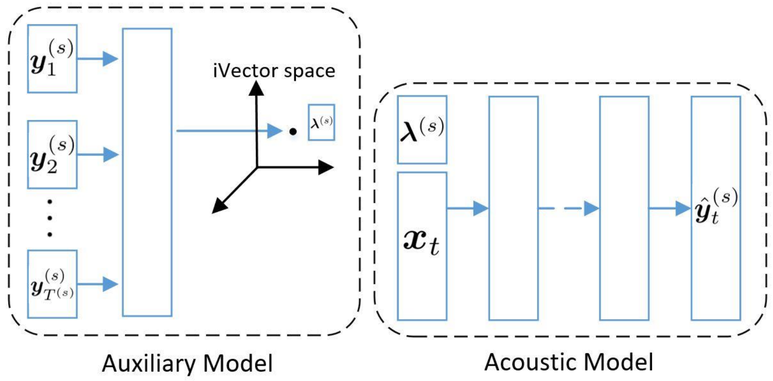
Vocoder

Contextual-Dependent Labels

State-Alignment

$y_1, y_2, y_3, \ldots y_T$ $x_1, x_2, x_3, \ldots, x_T$

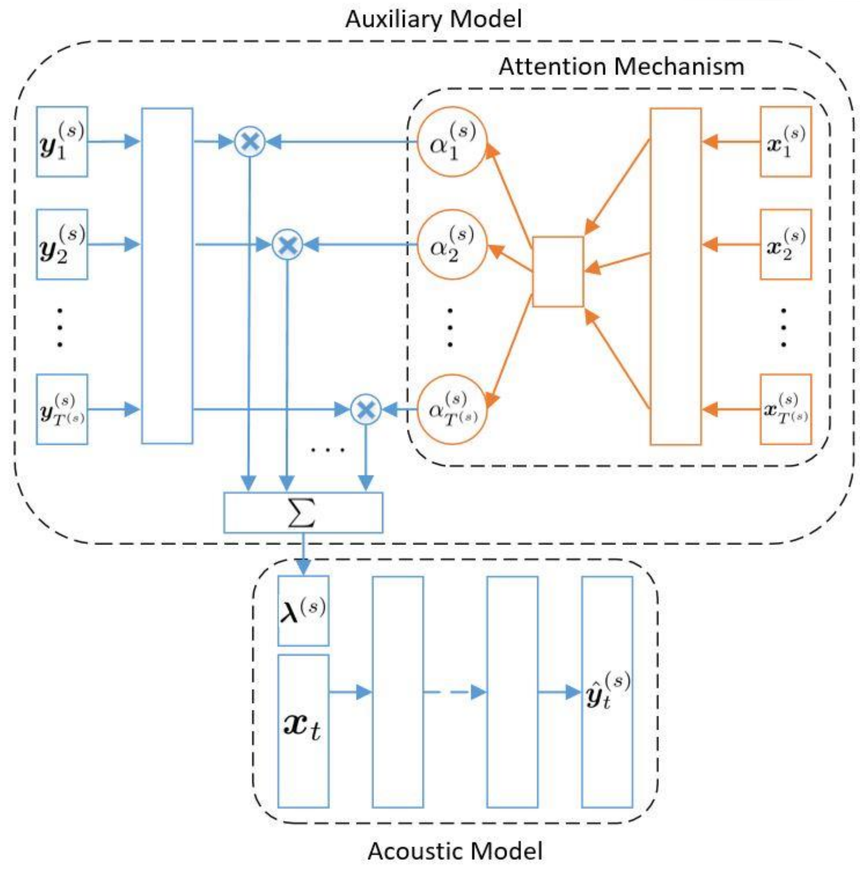
iVector-based Adaptive Speech Synthesis



- Learn a speaker space, each speaker corresponds to an iVector
- Previous iVectors: one-hot speaker code; d-vector, GMM-based i-vector

- **Problems:**
  - Independent and different optimisation criteria
  - Uniform weight on adaptive data
  - i-vector extraction from acoustic data only

Integrated iVector Extraction with Attention Mechanism (IIEA)



- Integrate the iVector extraction process in training
- Apply attention mechanism to adaptive data
- Learn attentions from contextual labels
- "Flat Attention" (**IIE**): remove the Attention Mechanism (top right) and use a uniform Attention of **1/T**.

Objective Measures

Model	MCD (dB)	BAP (dB)	F0 RMSE	F0 CORR	VUV error
baseline	6.703	2.229	30.463	**0.836**	5.924%
IIE	**6.601**	**2.220**	30.017	0.831	**5.770%**
IIEA	6.615	2.222	**29.889**	0.833	5.799%

Subjective Evaluations

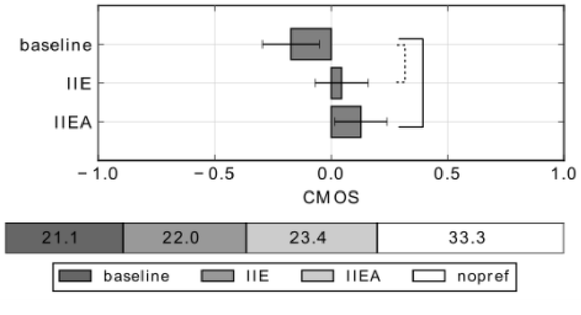


Figure 6: *naturalness preference test results, iVector size 128; 32 listeners took the test properly*

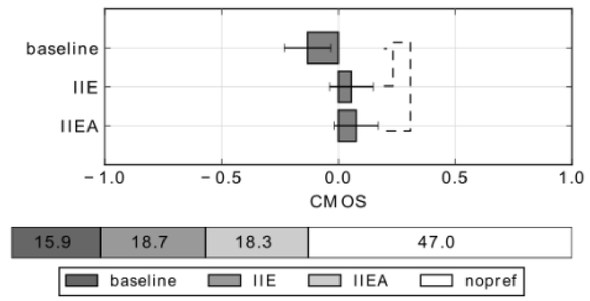
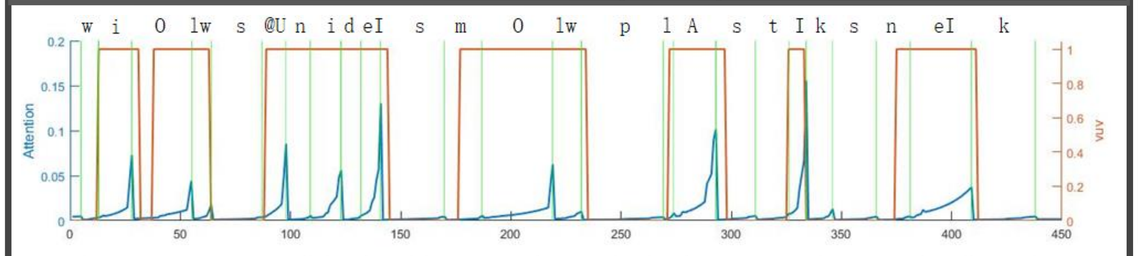


Figure 7: *similarity preference test results, iVector size 128; 33 listeners took the test properly*

Attention Trajectory



Phone sequence (black with green boundaries), VUV trajectory (red) and Attention trajectory (blue) : "We also need a small plastic snake"

Challenges

- Attention mechanism is hard to train
  - Initialisation: with a pre-trained IIE, bootstrap
  - Also initialise IIE with a pre-trained baseline
- Additional Depth, vanishing gradient
  - Hyper-parameter tuning
  - Balance of adaptive data and acoustic data

Conclusions

- Improved performance
- Reasonable Attention trajectory
  - Mostly in the voiced regions
  - Mostly in the vowel regions
  - More selective than a binary of either

[1] M. Wan, G. Degottex, and M. J. F. Gales, "Integrated speaker-adaptive speech synthesis." Accepted paper, ASRU, 2017