Minimum trajectory error training for deep neural networks, combined with stacked bottleneck features

Zhizheng Wu Simon King
The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

THE UNIVERSITY of EDINBURGH

Introduction

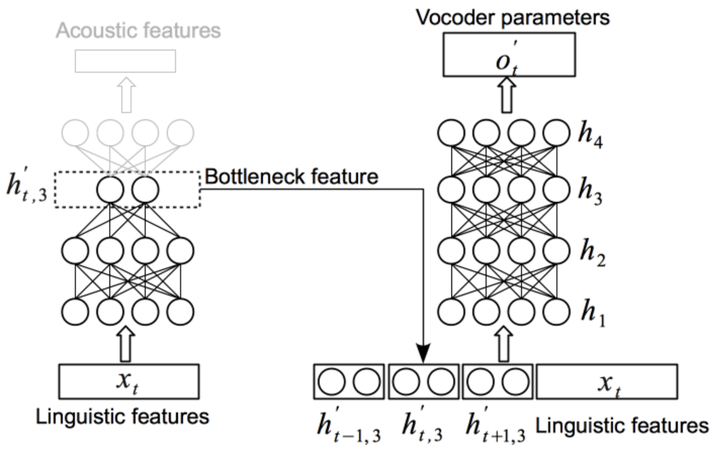
DNN-based speech synthesis performs frame-by-frame mapping without **contextual constraints** during training

Dynamic features are treated as stacked features during the training process

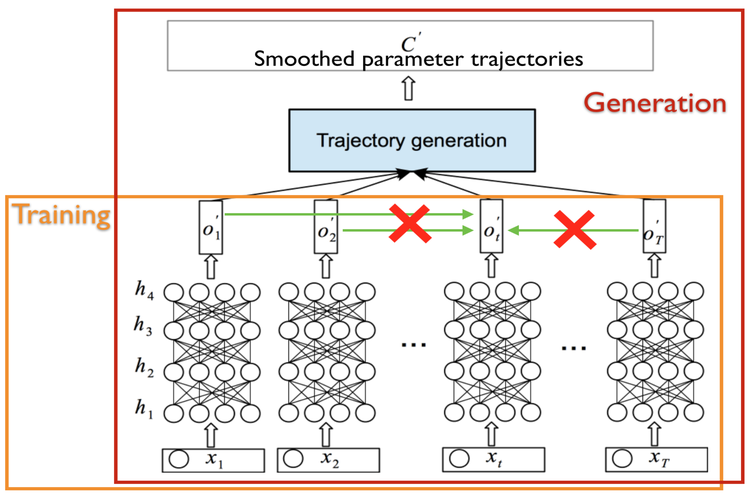
This work proposes a **minimum trajectory error training criterion** to minimise **utterance-level trajectory errors** rather than frame-by-frame errors

Background

Stacking bottleneck features for contextual constraints



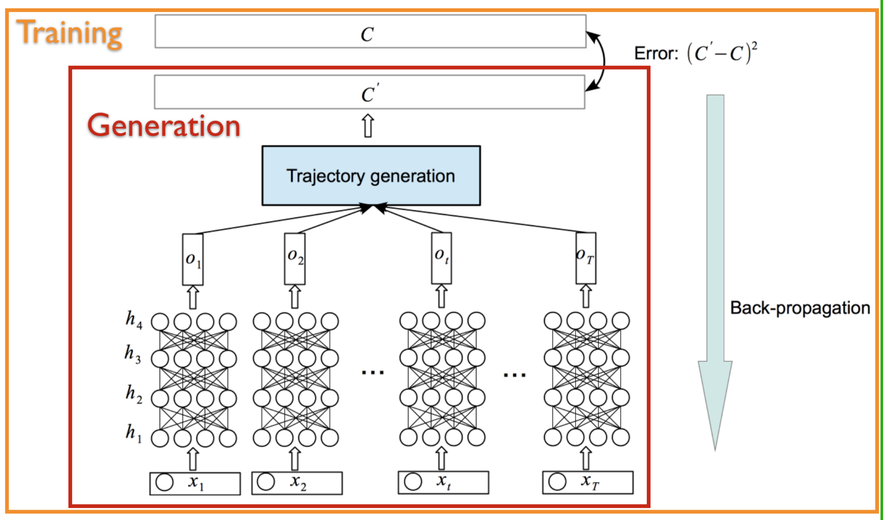
Minimising frame-by-frame errors during training, not taking dynamic constraints into consideration



$$D(\hat{\mathbf{o}}, \mathbf{o}) = (\hat{\mathbf{o}} - \mathbf{o})^\top (\hat{\mathbf{o}} - \mathbf{o})$$
$$\mathbf{o} = [\mathbf{c}^\top, \Delta\mathbf{c}^\top, \Delta^2\mathbf{c}^\top]^\top$$

Proposed minimum trajectory error training criterion

Miminise the **utterance-level vocoder parameter trajectory errors** rather than frame-by-frame errors



Trajectory generation: Maximum likelihood parameter generation (MLPG)

$$\hat{\mathbf{C}} = (\mathbf{W}^\top \mathbf{U}^{-1} \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{U}^{-1} \hat{\mathbf{O}}$$
$$\mathbf{R} = (\mathbf{W}^\top \mathbf{U}^{-1} \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{U}^{-1}$$
$$\mathbf{O} = \mathbf{W}\mathbf{C}$$

New objective function: **minimise the trajectory error after MLPG**, rather than before MLPG

$$D(\hat{\mathbf{C}}, \mathbf{C}) = (\hat{\mathbf{C}} - \mathbf{C})^\top (\hat{\mathbf{C}} - \mathbf{C}) = (\mathbf{R}\hat{\mathbf{O}} - \mathbf{C})^\top (\mathbf{R}\hat{\mathbf{O}} - \mathbf{C})$$

Experimental setup

Training: 2400, development: 70, testing: 72 --- sampling rate: 48 kHz

Vocoder parameters: from STRAIGHT. 60-D Mel-Cepstral coeffcients (MCC) with double deltas, 25-D Band Aperiodicity (BAP) and log-scale F_0

FE-DNN: feed-forward deep neural network using frame-by-frame mean squared error training criterion. six hidden layers, each layer has 1024 hidden units. Tangent activation function for hidden, linear activation function for output

MTE-DNN: DNN using the proposed minimum trajectory error training criterion

BN-DNN: DNN with stacked bottlenecks as input to include contextual constraints

MTE-BN-DNN: BN-DNN using the proposed MTE training criterions

Experimental results

Objective results:

MTE criterion can reduce objective measures, and the performance can be further improved after combining with stacked bottlenecks

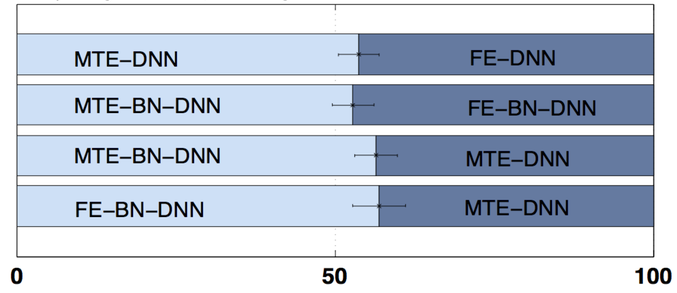
	MCD (dB)	F_0 RMSE (Hz)	V/UV error rate (%)
FE-DNN	4.19	9.13	4.24
MTE-DNN	4.12	8.93	4.28
FE-BN-DNN	4.03	8.91	3.97
MTE-BN-DNN	3.99	8.97	4.02

Subjective results:

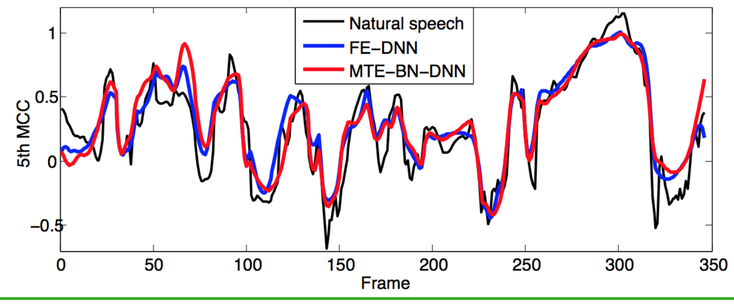
MTE criterion can **improve** the **naturalness** of synthesised speech significantly.

When MTE is combined with stacked bottlenecks, the improvement is not significant.

Comparing with MTE, stacking **bottleneck features** is **more effective**.



Comparison between the 5th MCC trajectories from natural, FE-DNN and MTE-BN-DNN
The trajectory from MTE-BN-DNN is more close to the natural one.



Conclusions

The proposed minimum trajectory error training criterion can improve the naturalness of synthesised speech significiantly.

The new training criterion can be integrated with stacked bottleneck to further improve the performance

References

1. Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, Simon King, "Deep neural network employing multi-task learning and stacked bottleneck features for speech synthesis", ICASSP 2015
2. Yi-Jian Wu, and Ren-Hua Wang. "Minimum generation error training for HMM-based speech synthesis." ICASSP 2006
3. Tokuda, Keiichi, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. "Speech parameter generation algorithms for HMM-based speech synthesis." ICASSP 2000