Deep neural network context embeddings for model selection in rich-context HMM synthesis

Thomas Merritt, Junichi Yamagishi, Zhizheng Wu, Oliver Watts, Simon King
T.Merritt@ed.ac.uk
Centre for Speech Technology Research, University of Edinburgh
National Institute of Informatics, Japan

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

THE UNIVERSITY *of* EDINBURGH

CSTR

Abstract

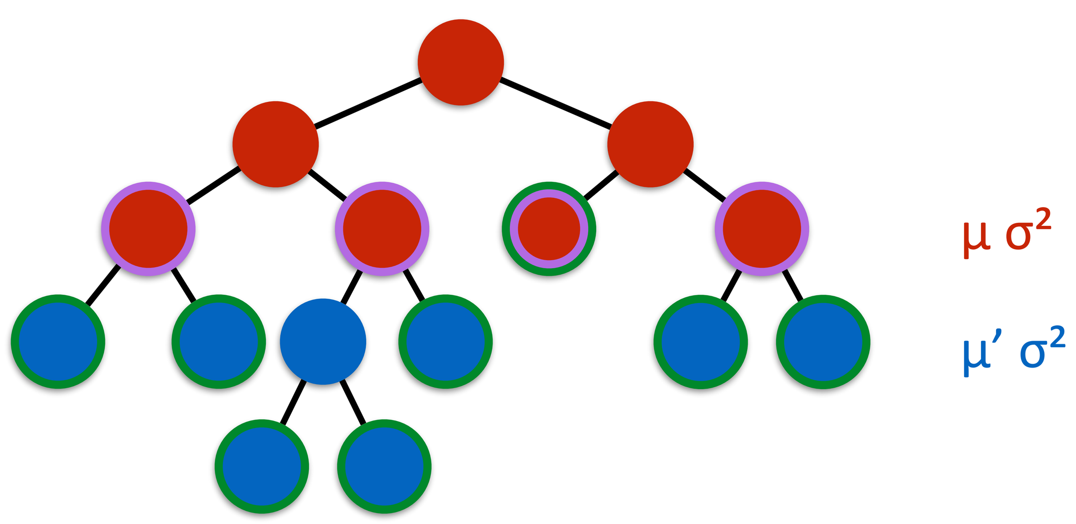
Background:

- Across-linguistic context averaging is extremely detrimental to quality. Within-linguistic context averaging is much more preferable [1].
- Conventional rich-context synthesis system [2] - modelling within-linguistic contexts only.

Contribution:

- Bottleneck features extracted using [3] are used to identify closest rich-context models where out-of-training contexts are encountered.

Rich-context models



- Conventional trained decision tree
- Untie leaf nodes
- Update means (keep tied variances)

Proposed rich-context selection

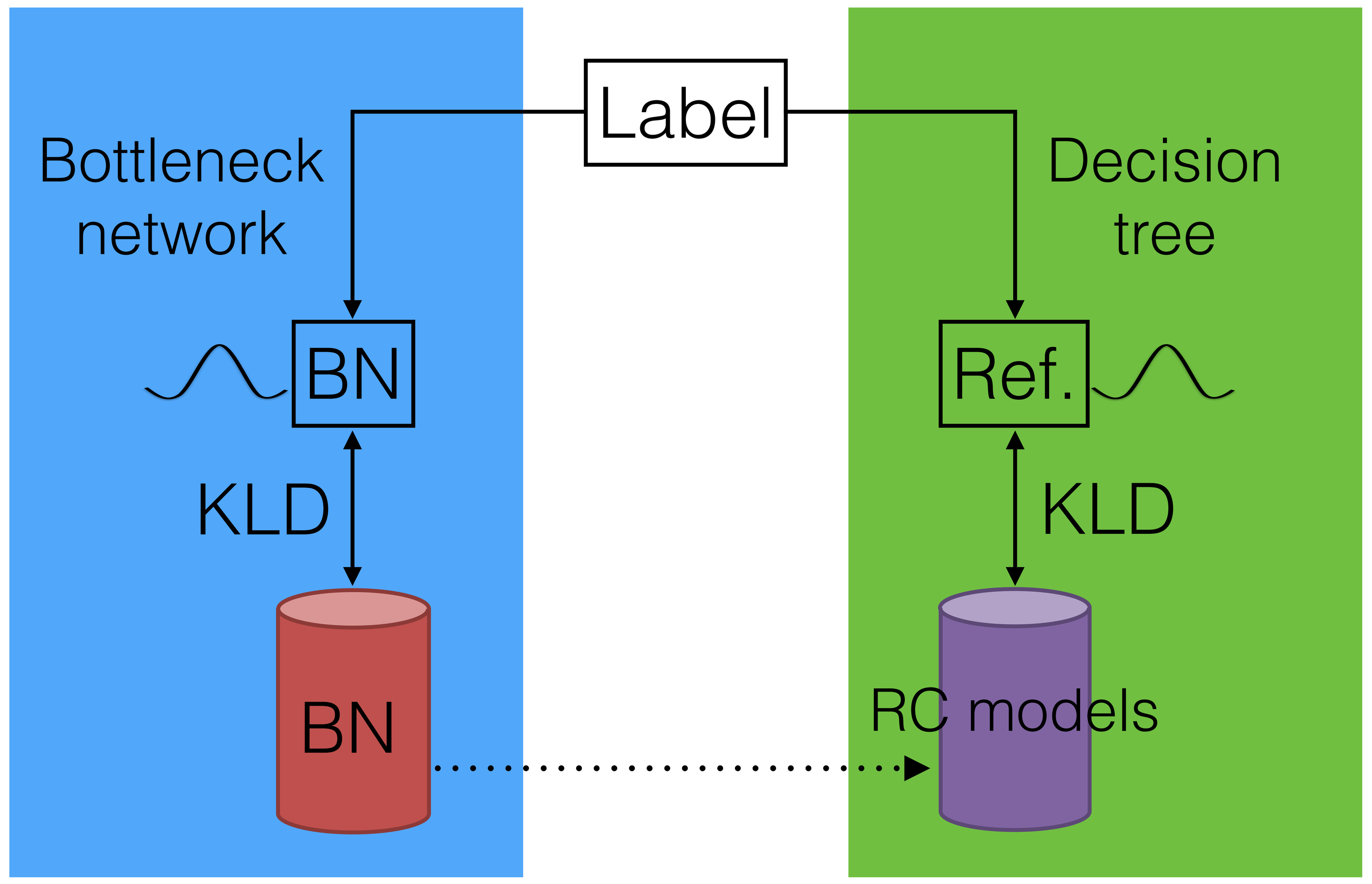


Figure: **Blue:** proposed system. **Green:** conventional rich-context system.

At training time

- Frame-wise bottleneck features (BN) generated using [3]
- HMM state alignments used
- Distributions of training context features calculated

At synthesis time

- Frame-wise bottleneck features generated
- HMM state alignments used
- Closest seen rich-context (RC) model selected based on distance in 'bottleneck space'
- For each phoneme distances across all states summed together to guide selection

System comparisons

Standard HMM system

- Select tied cluster from decision tree. Calculated using across-context averaging - reduces quality [1].

Conventional rich-context system

- Pre-selection of rich contexts to use based on matching triphone.
- Kullback-Leibler divergence (KLD) between **standard tied cluster** & each rich context.
- Smallest divergence selected.

Proposed system

- No linguistic constraints placed - this is learnt by DNN.

Results

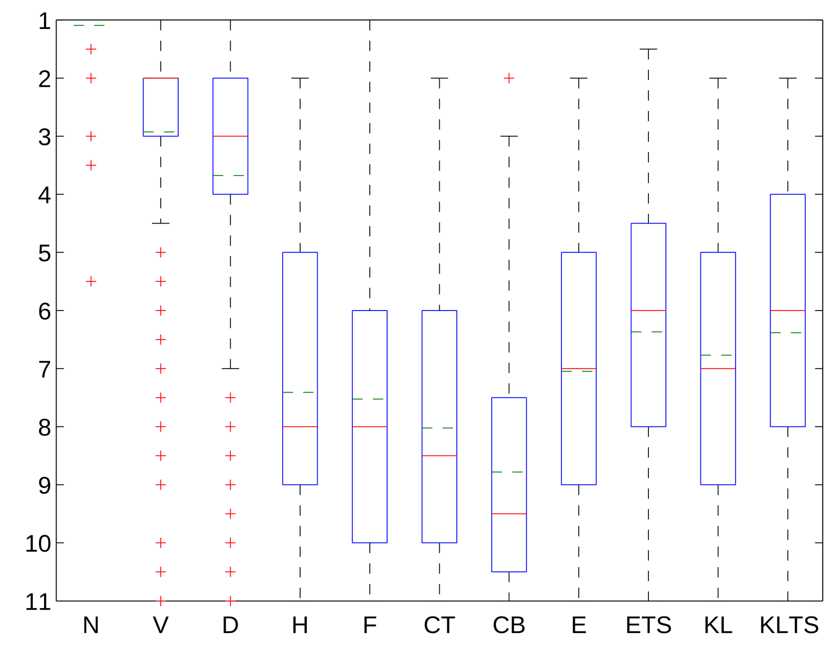


Figure: Boxplot of rank order of conditions from MUSHRA test

- N - natural
- V - vocoded
- D - Stacked bottleneck DNN [3]
- H - HTS demo
- F - fully untied tree (MDL=0)
- CT - [2] w/ triphone pre-selection
- CB - [2] w/ biphone pre-selection
- E - proposed - Euclidean distance
- ETS - proposed - Euclidean distance w/ tied source
- KL - proposed - KLD
- KLTS - proposed - KLD w/ tied source

Conclusions & future work

- Proposed system provides significantly improved selection of rich-context models.
- Pre-selection in [2] inadvertently hiding that target distribution is not optimal.
- DNN system no longer requires speech parameters as output - perceptually more relevant features can be used.

References

[1] T. Merritt, J. Latorre and S. King. Attributing modelling errors in HMM synthesis be stepping gradually from natural to modelled speech. ICASSP 2015

[2] Zhi-jie Yan, Yao Qian and F.K. Soong. Rich context modeling for high quality HMM-based TTS. Interspeech 2009

[3] Z. Wu, C. Valentini-Botinhao, O. Watts and S. King. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. ICASSP 2015