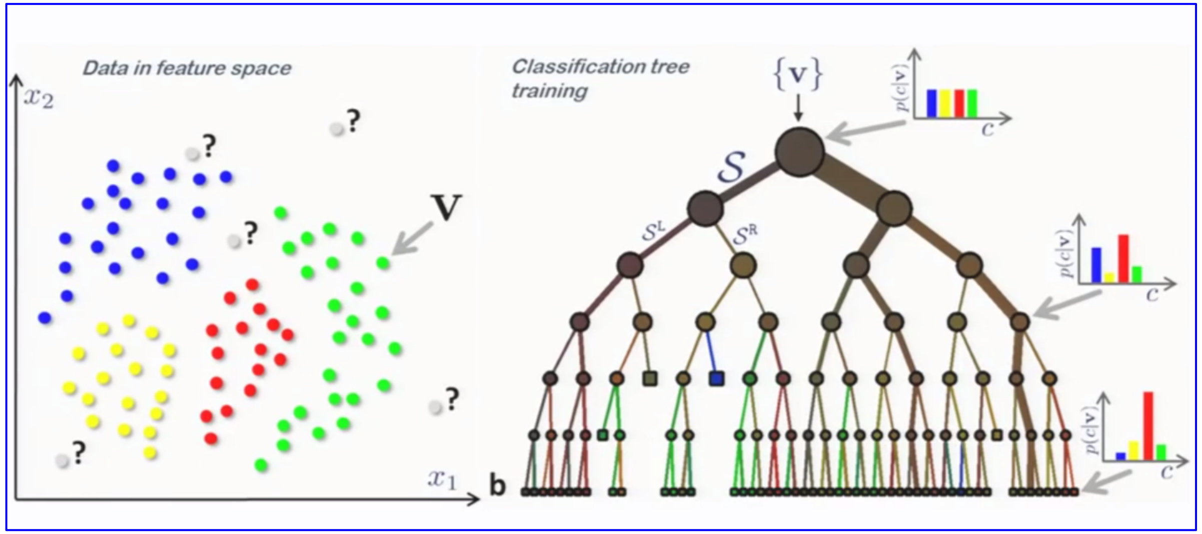
Summary of Speech Synthesis sessions at Interspeech 2015

Sai Krishna
Speech and Vision Laboratory,
IIIT- Hyderabad

Random Forests for Speech Synthesis - 1

Decision Trees:



Idea is to find the attribute that lowers the amount of information required to completely describe each data point.

a. Choose the next best node.
b. Create a new decision tree at the chosen node.

Fig 1: Overview of Decision Tree based clustering

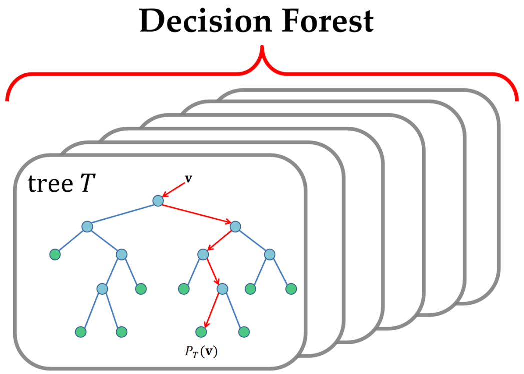
Disadvantages:

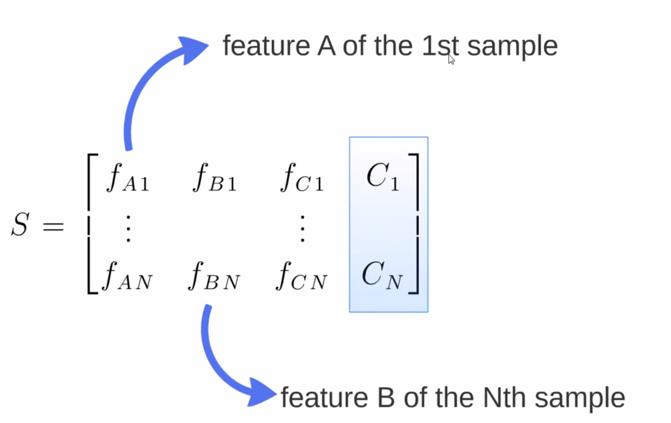
1. Over-split the data - reducing the number of examples too fast.
2. May not obtain the best result.

Random Forests for Speech Synthesis - 2

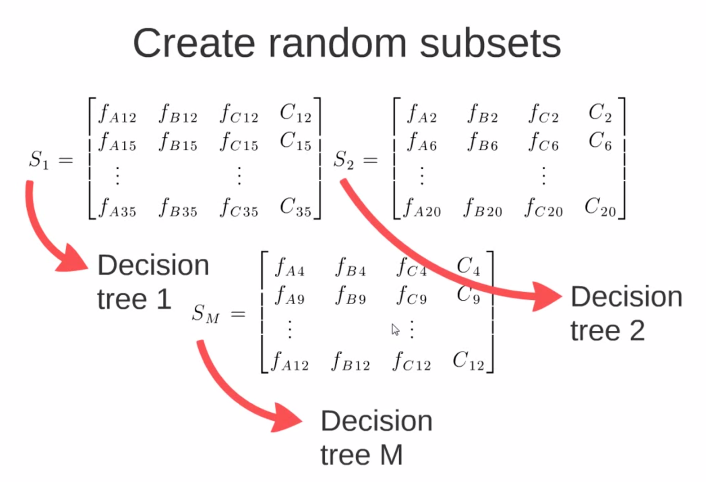
The motivation is that a combination of learning models is better compared to a single model.

Inherent advantage of avoiding oversplitting of the data, hence suitable when less data is available.





$$S = \begin{bmatrix} f_{A1} & f_{B1} & f_{C1} & C_1 \\ \vdots & & \vdots & \\ f_{AN} & f_{BN} & f_{CN} & C_N \end{bmatrix}$$



Create random subsets

$$S_1 = \begin{bmatrix} f_{A12} & f_{B12} & f_{C12} & C_{12} \\ f_{A15} & f_{B15} & f_{C15} & C_{15} \\ \vdots & & \vdots & \\ f_{A35} & f_{B35} & f_{C35} & C_{35} \end{bmatrix} \quad S_2 = \begin{bmatrix} f_{A2} & f_{B2} & f_{C2} & C_2 \\ f_{A6} & f_{B6} & f_{C6} & C_6 \\ \vdots & & \vdots & \\ f_{A20} & f_{B20} & f_{C20} & C_{20} \end{bmatrix}$$

$$S_M = \begin{bmatrix} f_{A4} & f_{B4} & f_{C4} & C_4 \\ f_{A9} & f_{B9} & f_{C9} & C_9 \\ \vdots & & \vdots & \\ f_{A12} & f_{B12} & f_{C12} & C_{12} \end{bmatrix}$$

Random Forests for Speech Synthesis - 3

Effect of adding Trees

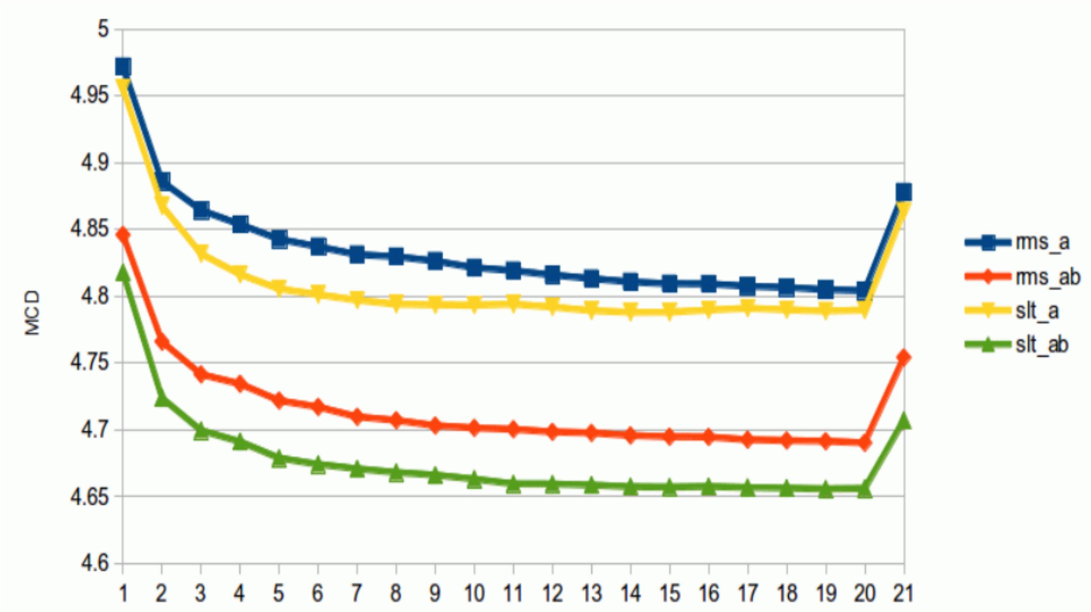


Fig:2 Effect of adding trees to the model

50% probability of picking each of the 63 linguistic features.

Effect of ignoring features

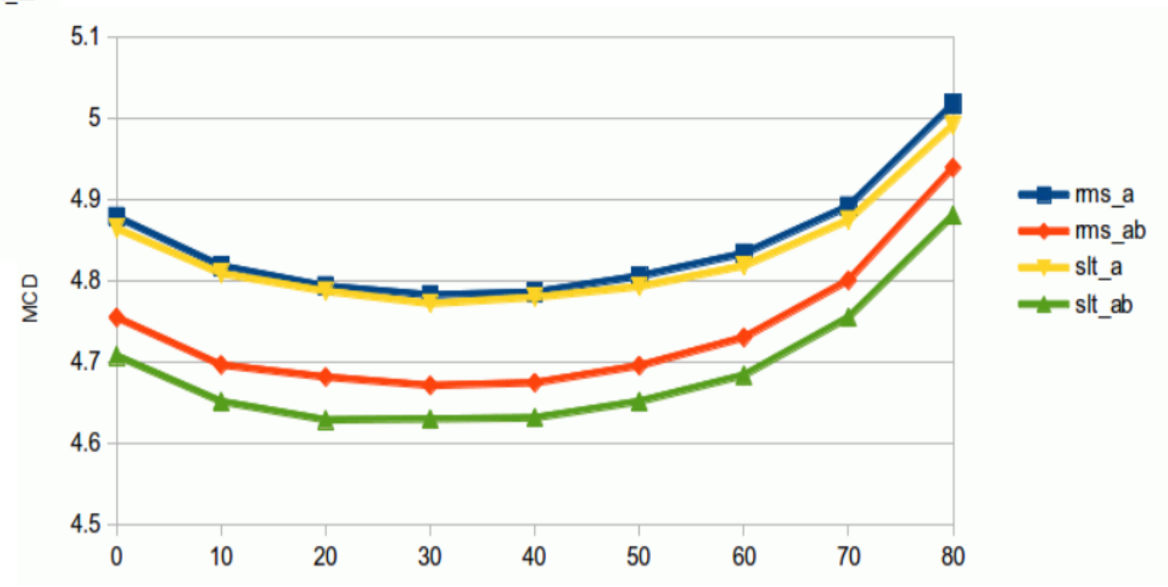


Fig 3: Effect of ignoring features

Duration Prediction using Multi Level Model for GPR-based Speech Synthesis

Issues in Modeling Duration:

Factor Confounding.

Data Sparsity.

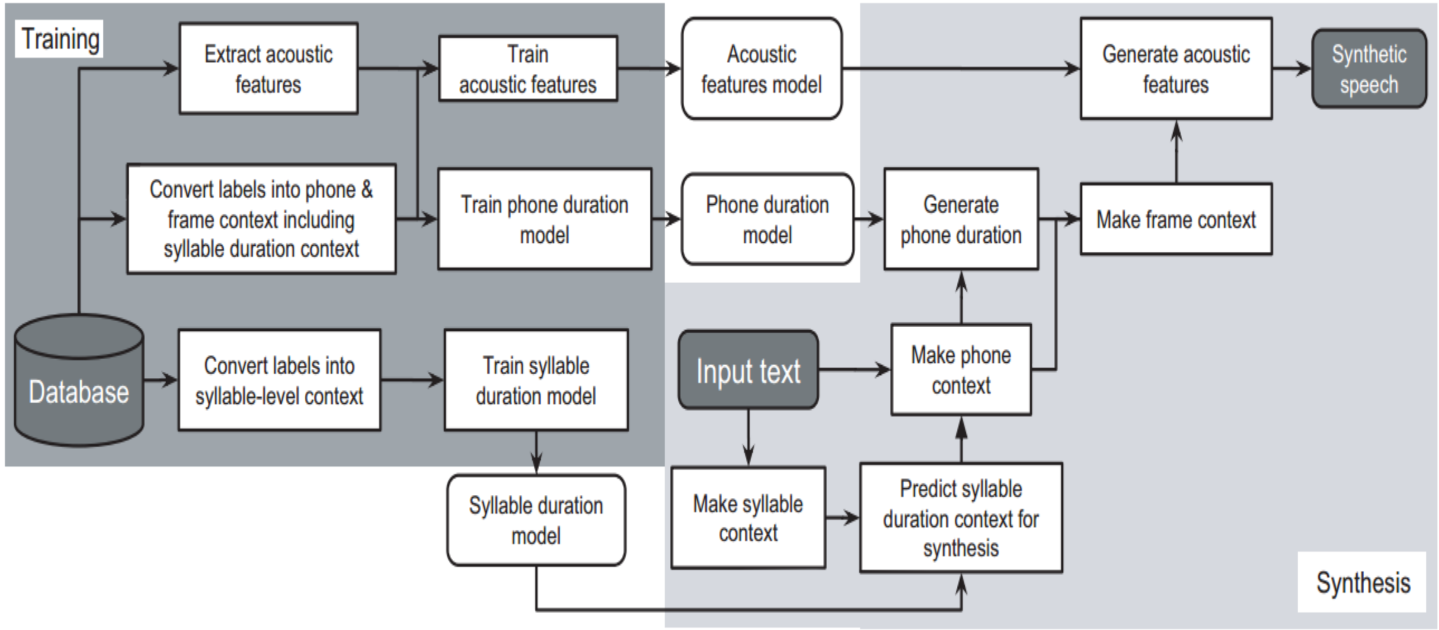

Training
Extract acoustic features
Train acoustic features
Acoustic features model
Generate acoustic features
Synthetic speech
Convert labels into phone & frame context including syllable duration context
Train phone duration model
Phone duration model
Generate phone duration
Make frame context
Database
Convert labels into syllable-level context
Train syllable duration model
Input text
Make phone context
Syllable duration model
Make syllable context
Predict syllable duration context for synthesis
Synthesis

Fig4 : GPR based Synthesis based on multi level duration prediction

Sentence level control vectors for deep neural network speech synthesis

An unsupervised manner of space of sentences which captures the dimensions of variation in the

training data.

Can be used to modulate the characteristics of synthetic speech on a sentence by sentence basis.

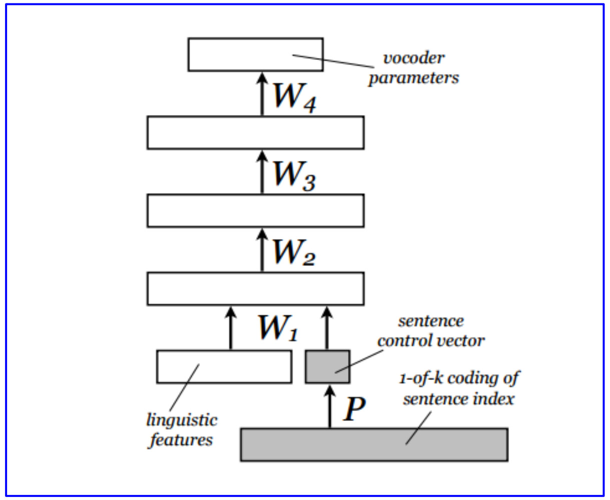


Fig 5 : Overview of the training process

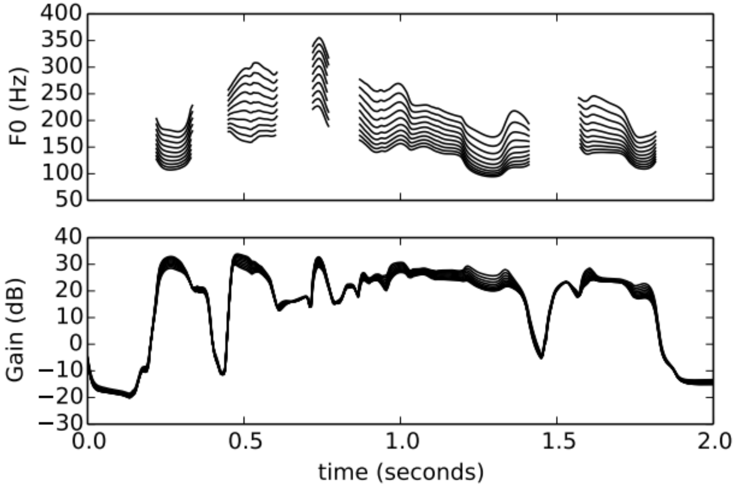


Fig 6 : Variation in synthetic f0 and gain

Using Deep Bidirectional Recurrent Neural Networks for Prosodic-Target Prediction in a Unit-Selection Text-to-Speech System

1. Explore the use of Bidirectional RNNs within unit selection synthesis systems.
2. Investigate their effect on unit search and evaluate it against the decision tree based approach.

a. Allowing for signal manipulation, can BiRNN models outperform baseline systems in perceptual evaluation.

b. What is the effect of the unit search of this improved prosodic target contour.

c. How do the baseline and BiRNN approaches compare when prosodic targets are used to drive the unit selection search and then natural prosody of the units is favoured in the output waveform.

Pruning Redundant Synthesis units based on Static and Delta Unit Appearance Frequency

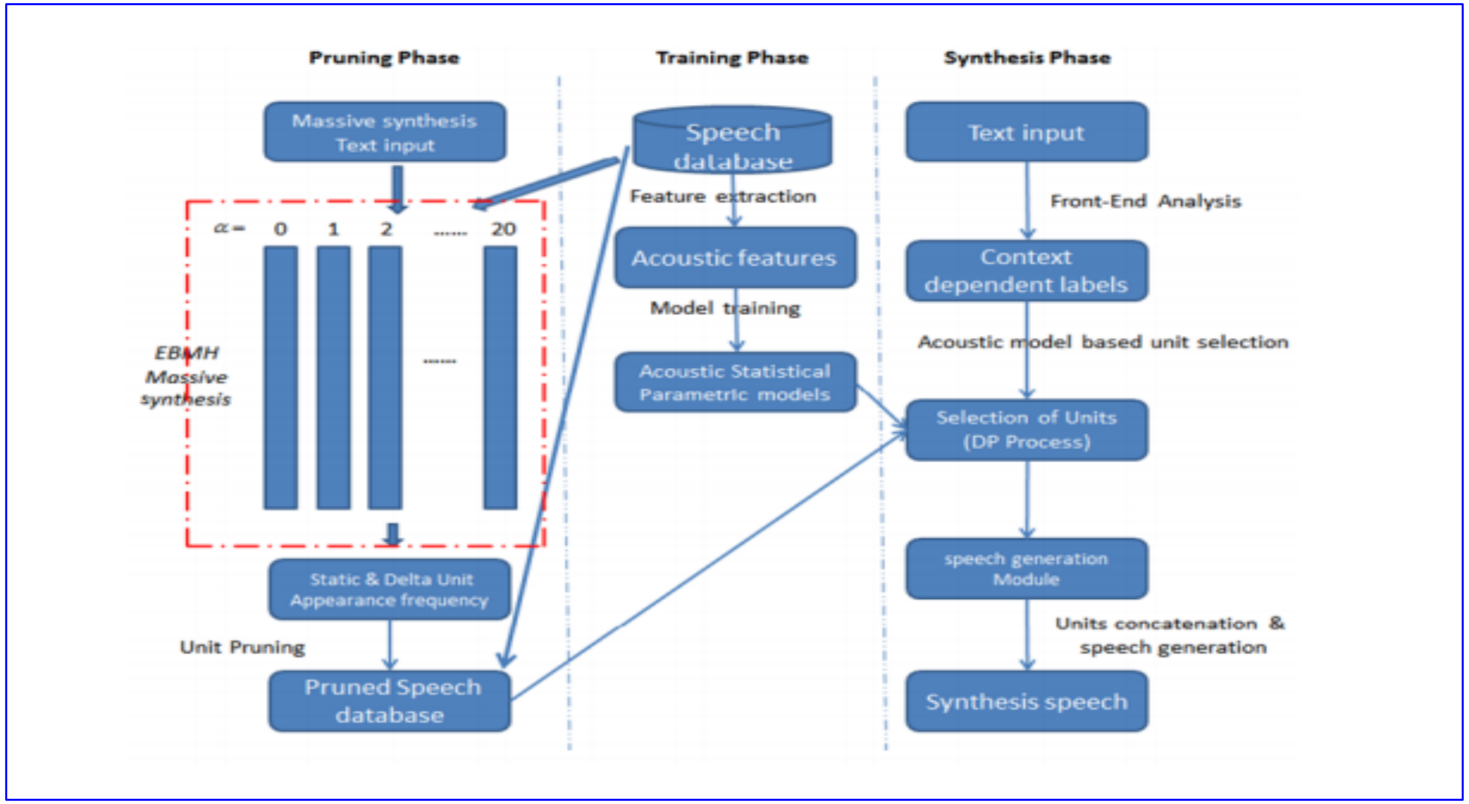


Fig 7 : Proposed Data pruning Technique