A template-based approach for speech synthesis intonation generation using LSTMs

Srikanth Ronanki

Gustav

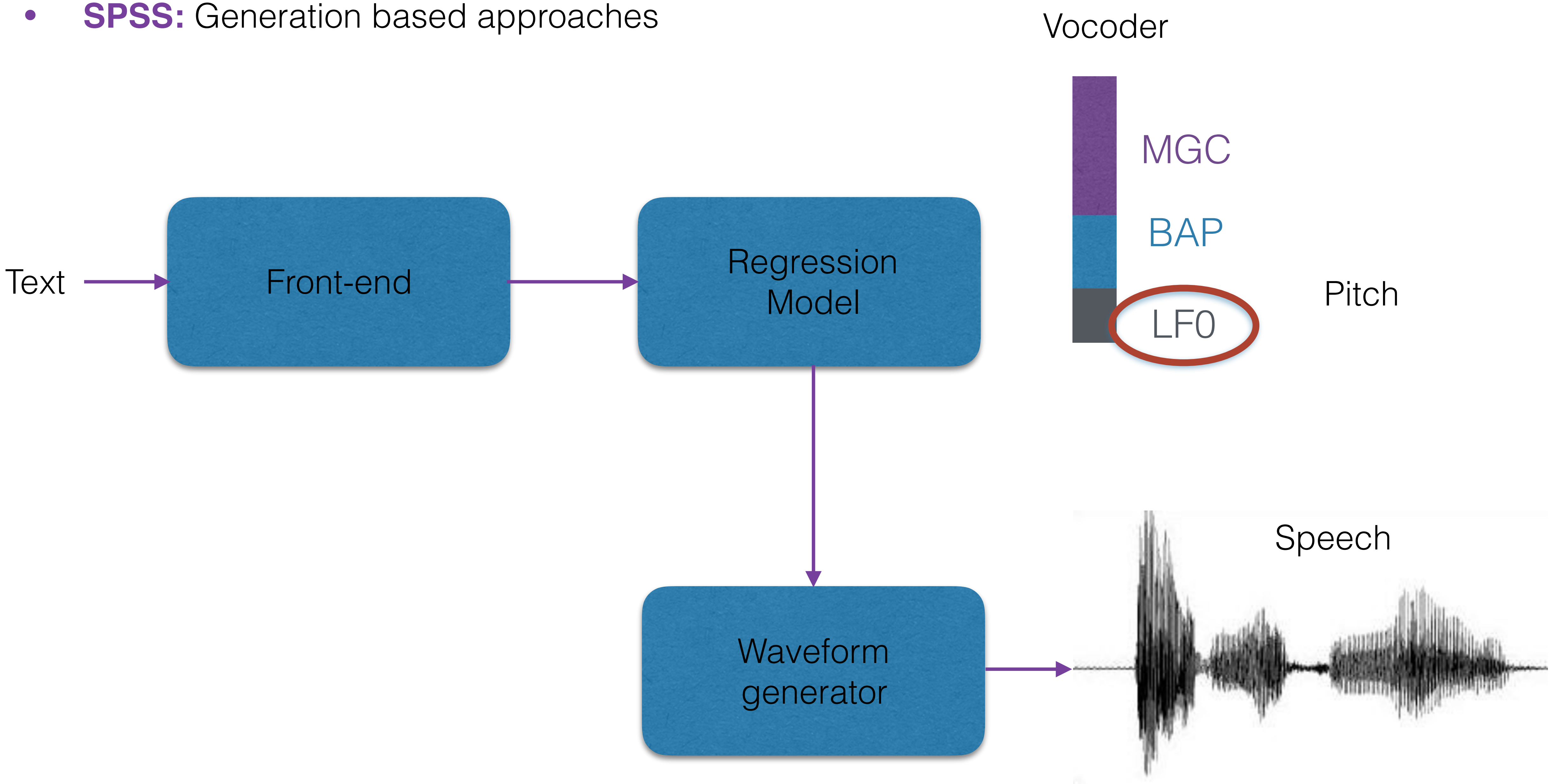
Zhizheng

Simon

THE UNIVERSITY *of* EDINBURGH
informatics

Introduction: Statistical speech synthesis

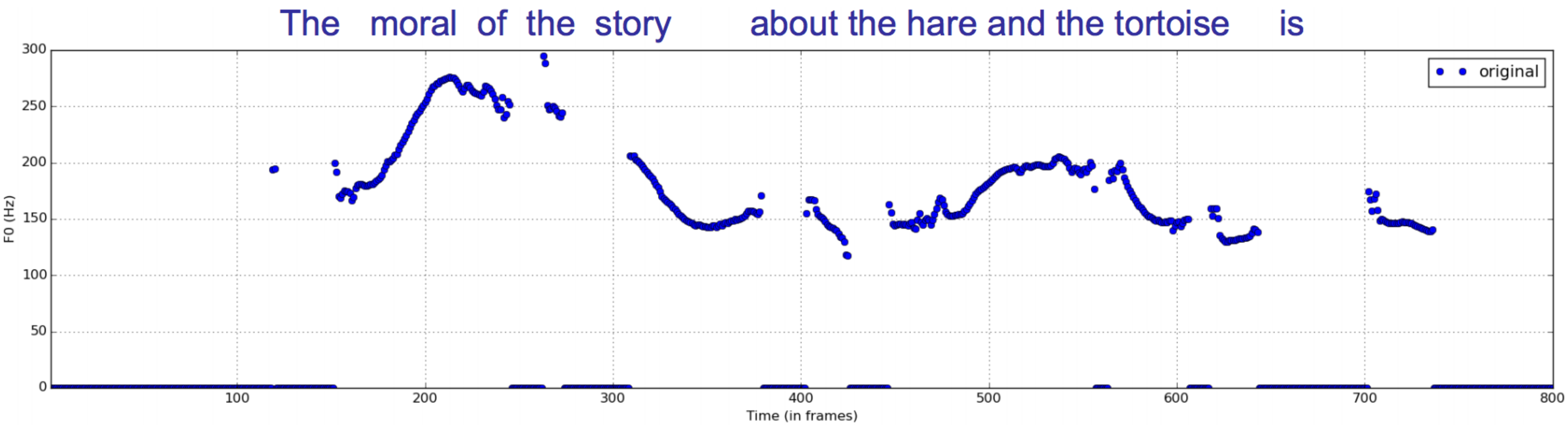
- **SPSS:** Generation based approaches

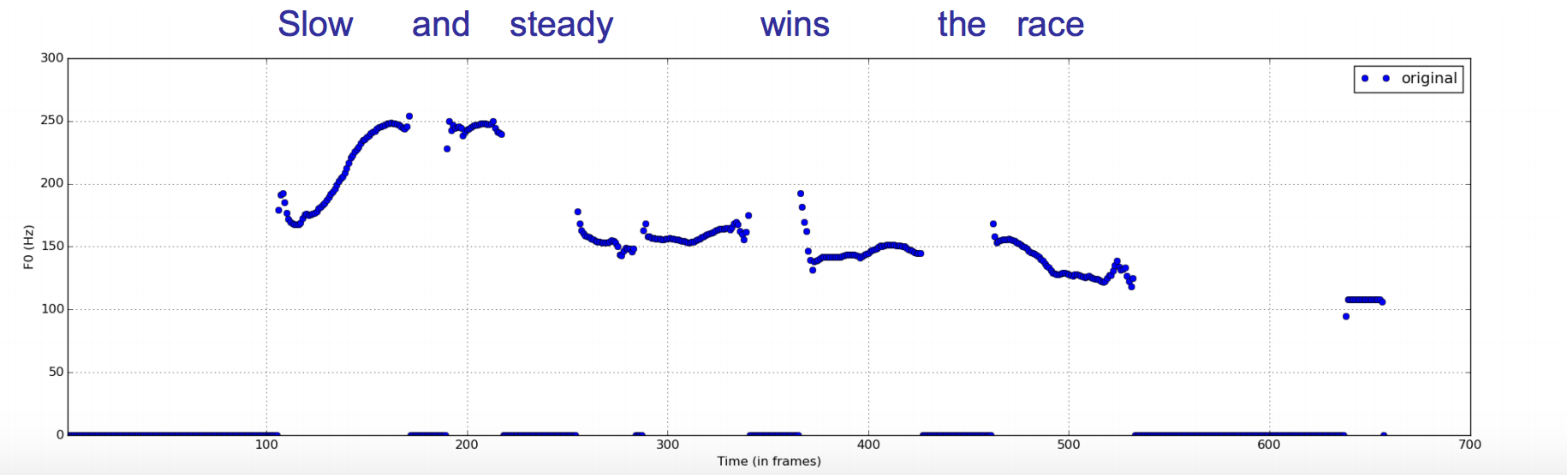


Why template-based approach?

- Lack of convincing intonation makes current parametric systems sound dull and lifeless.
- Typically, these systems predict F0 frame-by-frame using regression models.
- This approach leads to overly-smooth pitch contours and fail to construct an appropriate prosodic structure.
- Templates retain the dynamic range of F0 within the segment.
- We propose a classification-based approach to automatic F0 generation.

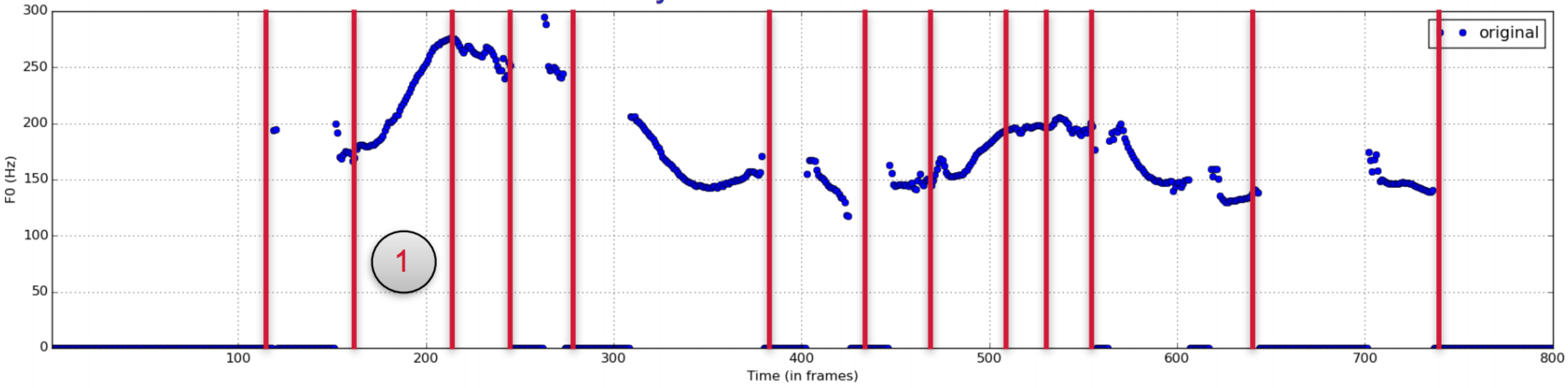
Pitch contour



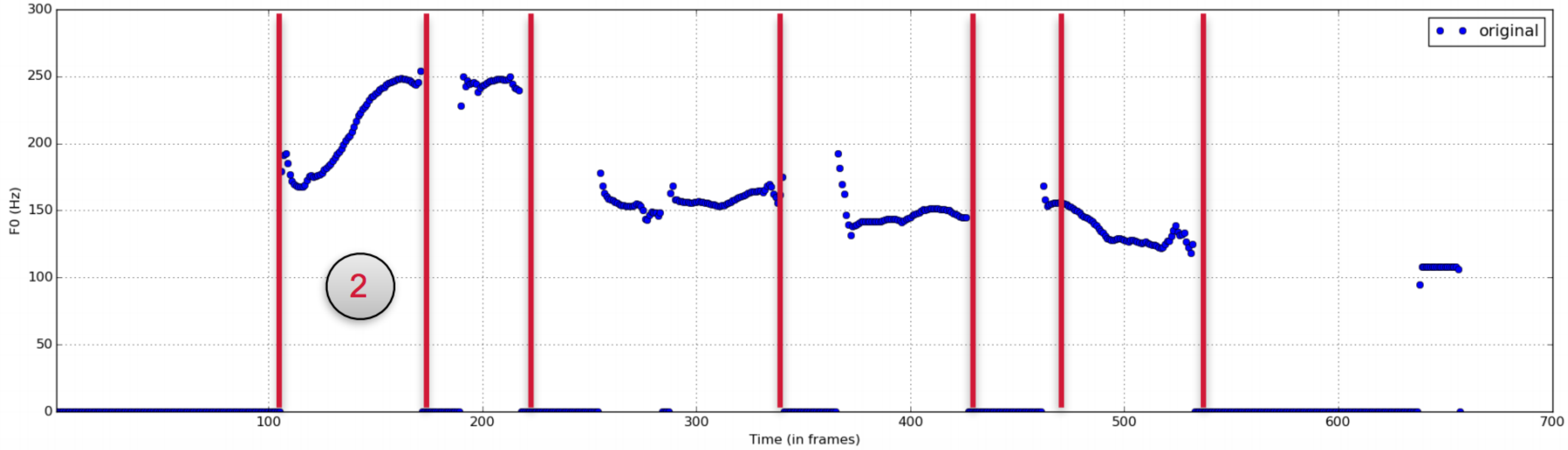


Pitch contour segmentation

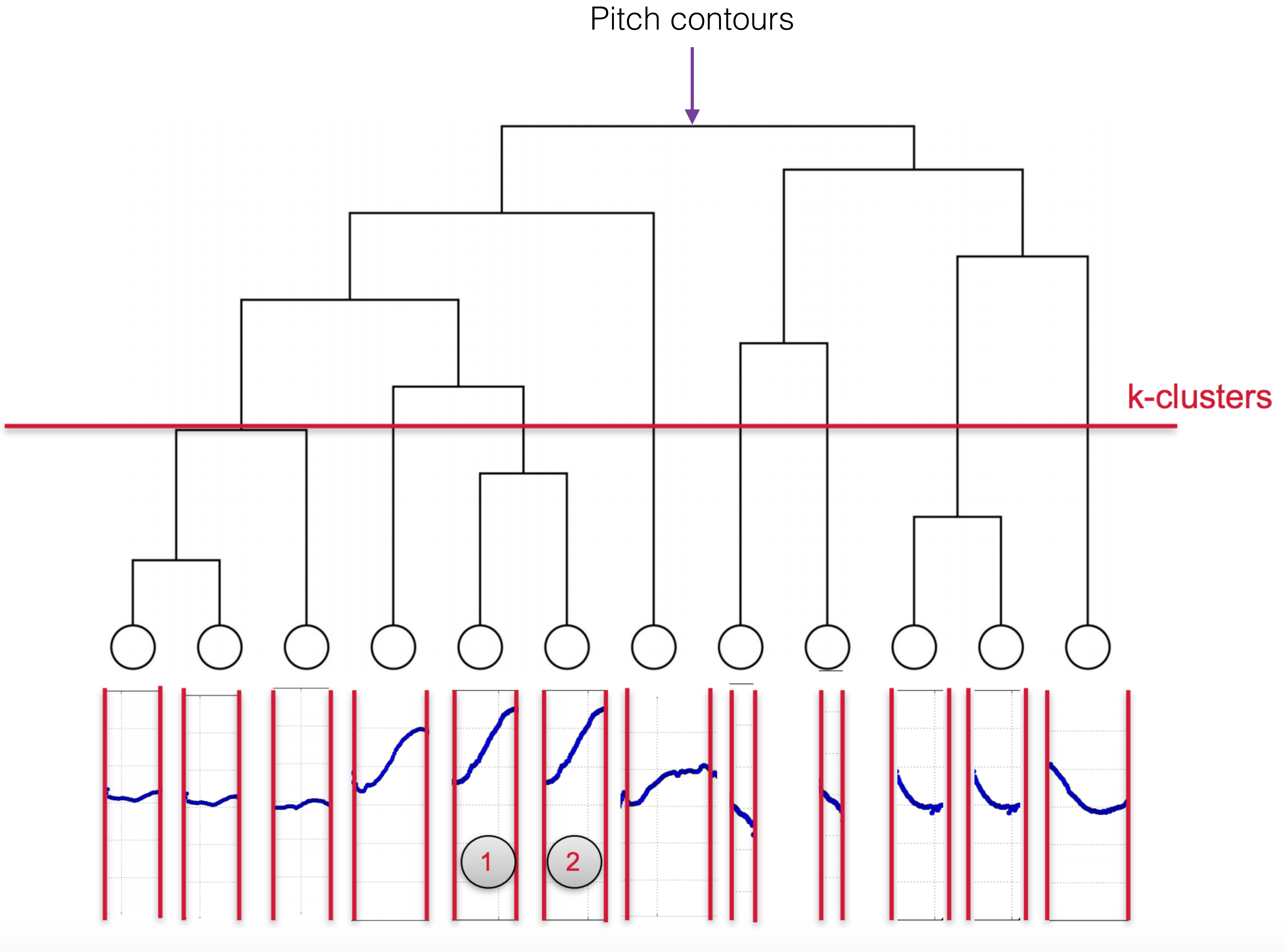
The moral of the story about the hare and the tortoise is



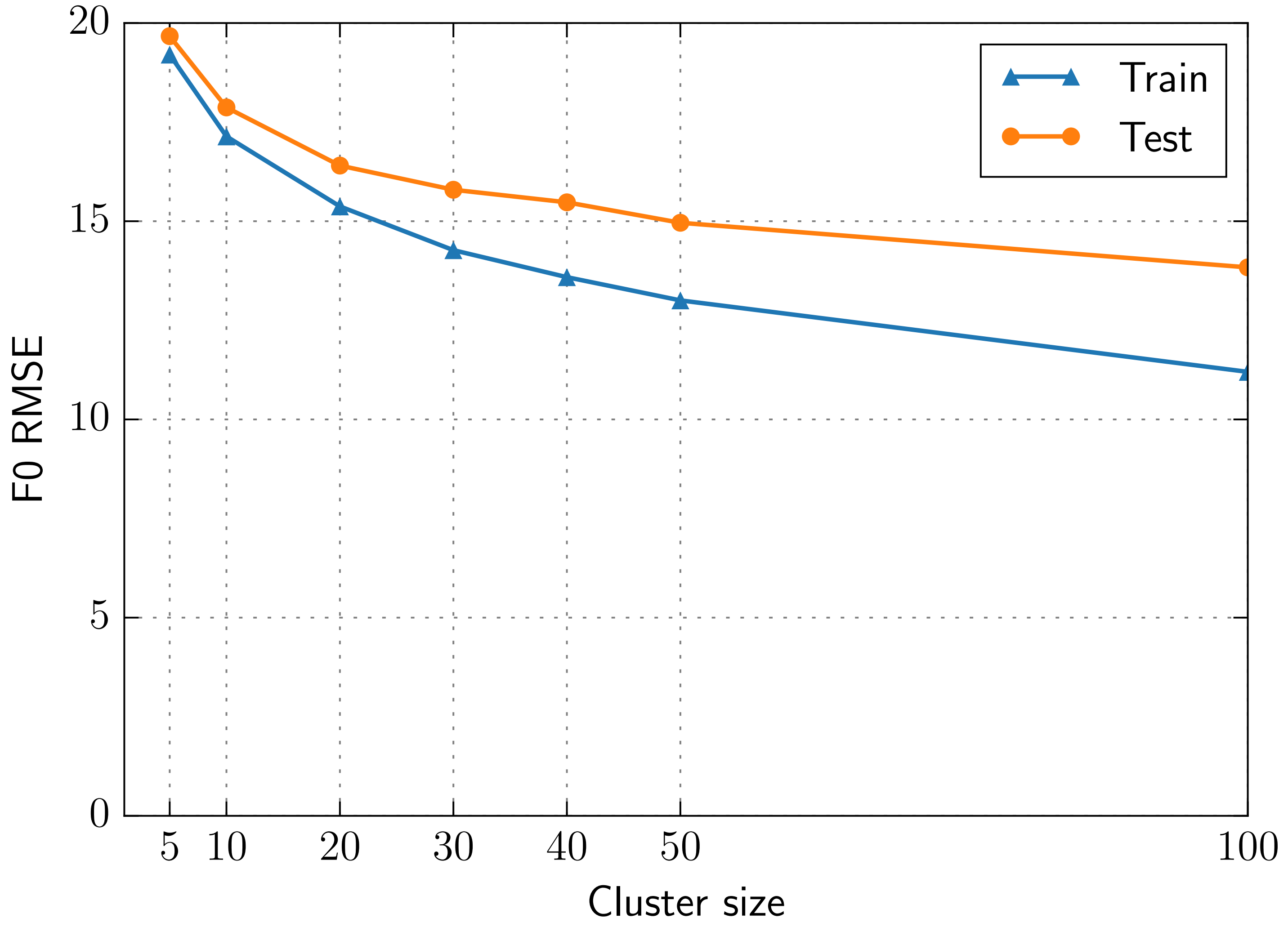
Slow and steady wins the race



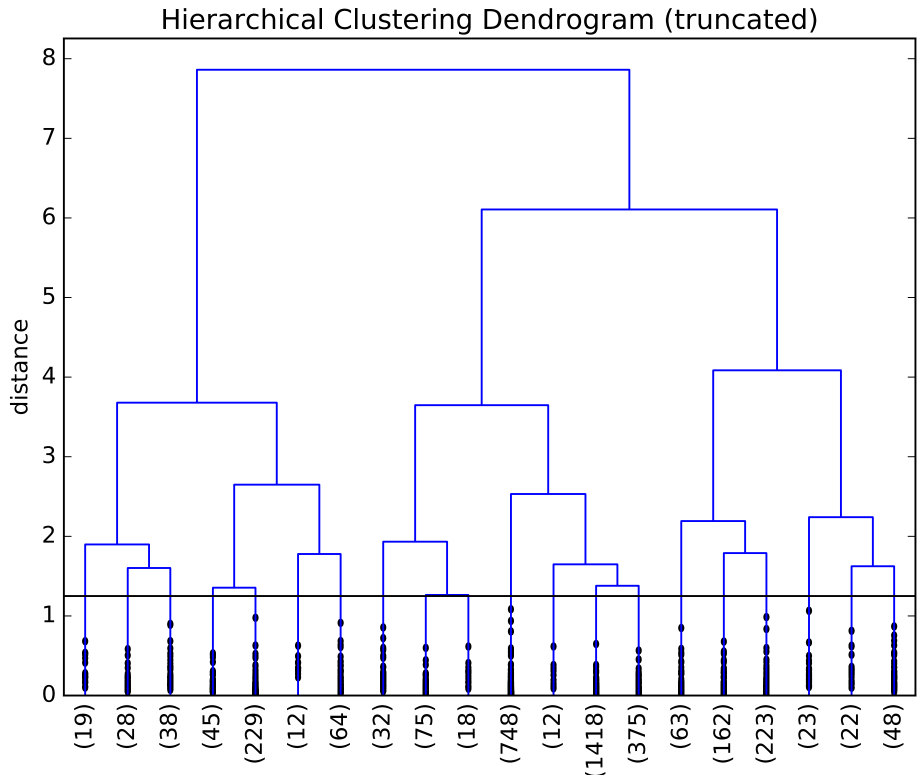
Hierarchical clustering

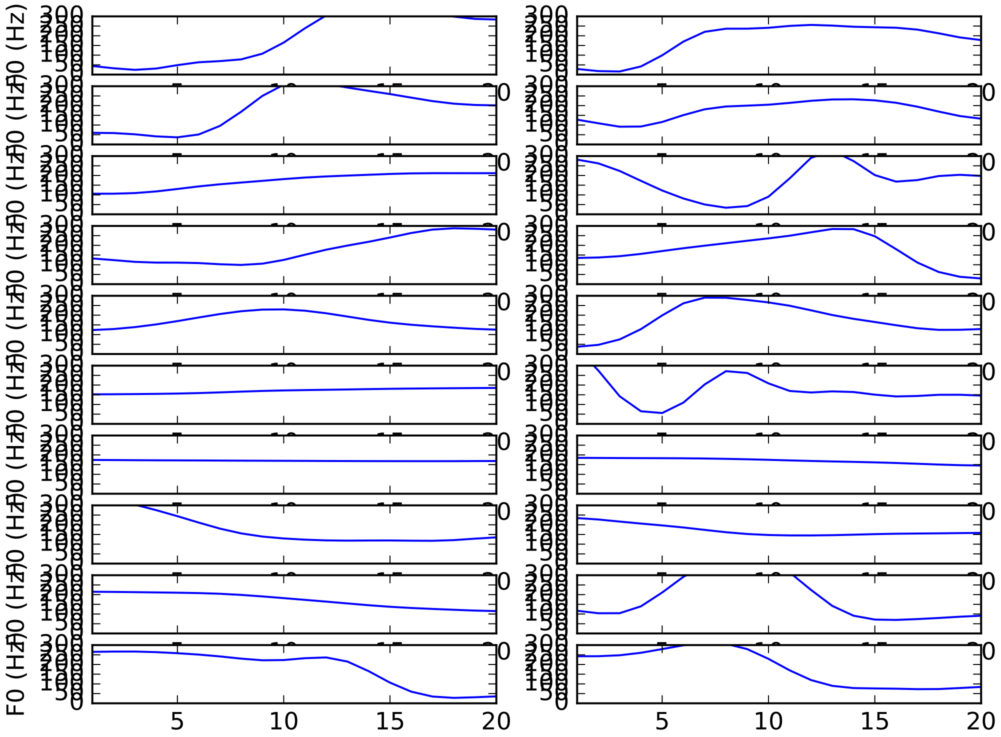


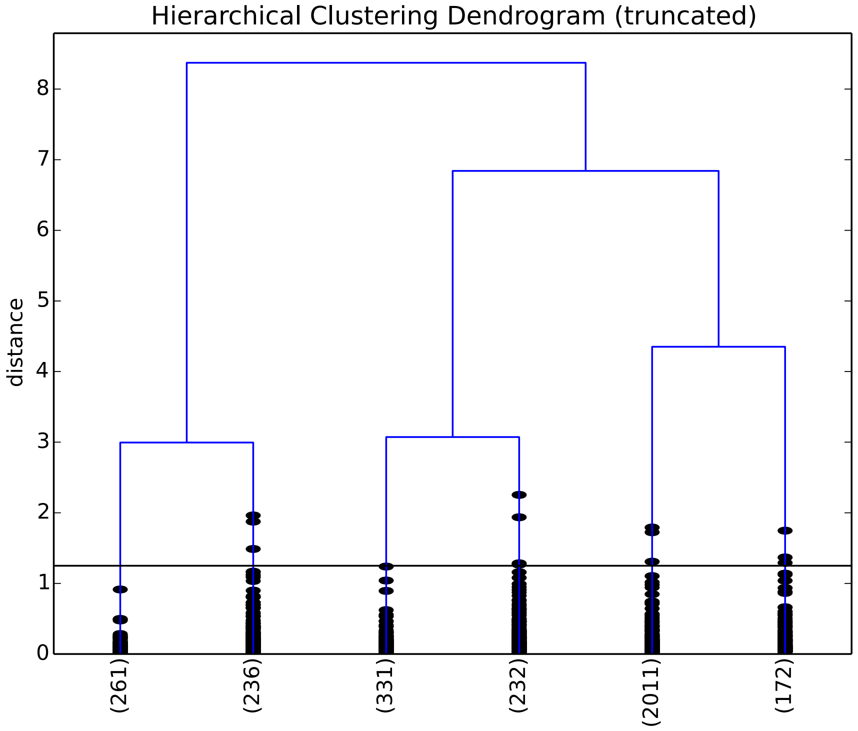
How to determine number of clusters?

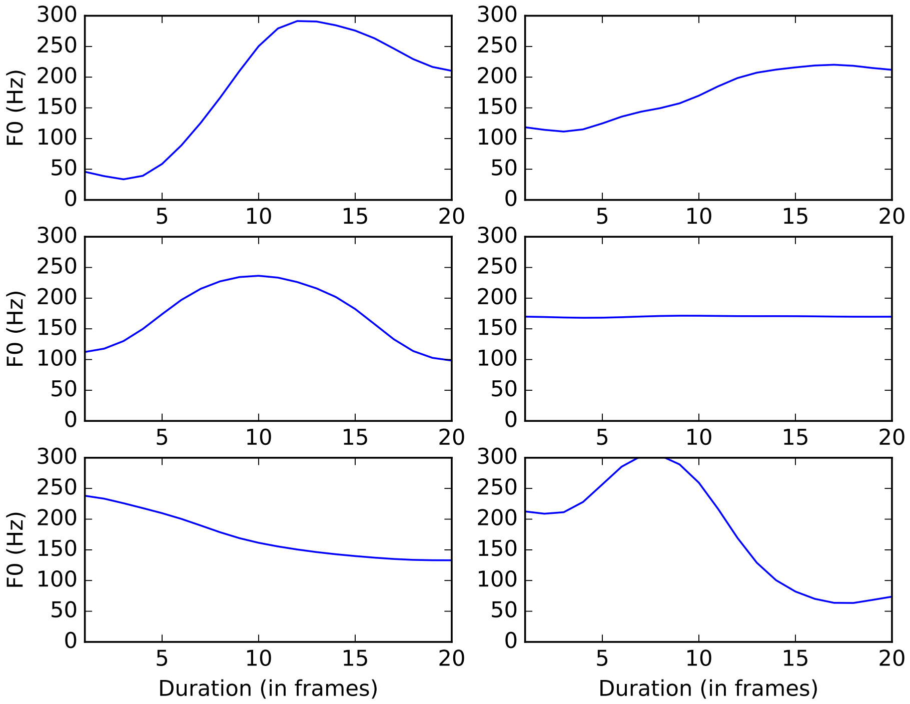


A set of templates (clusters)





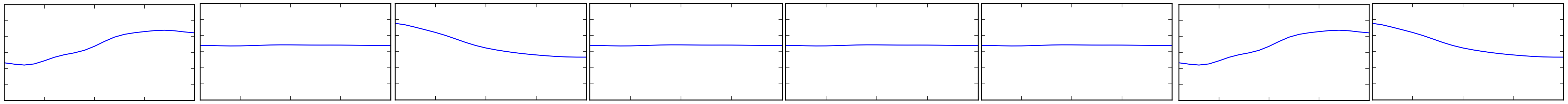


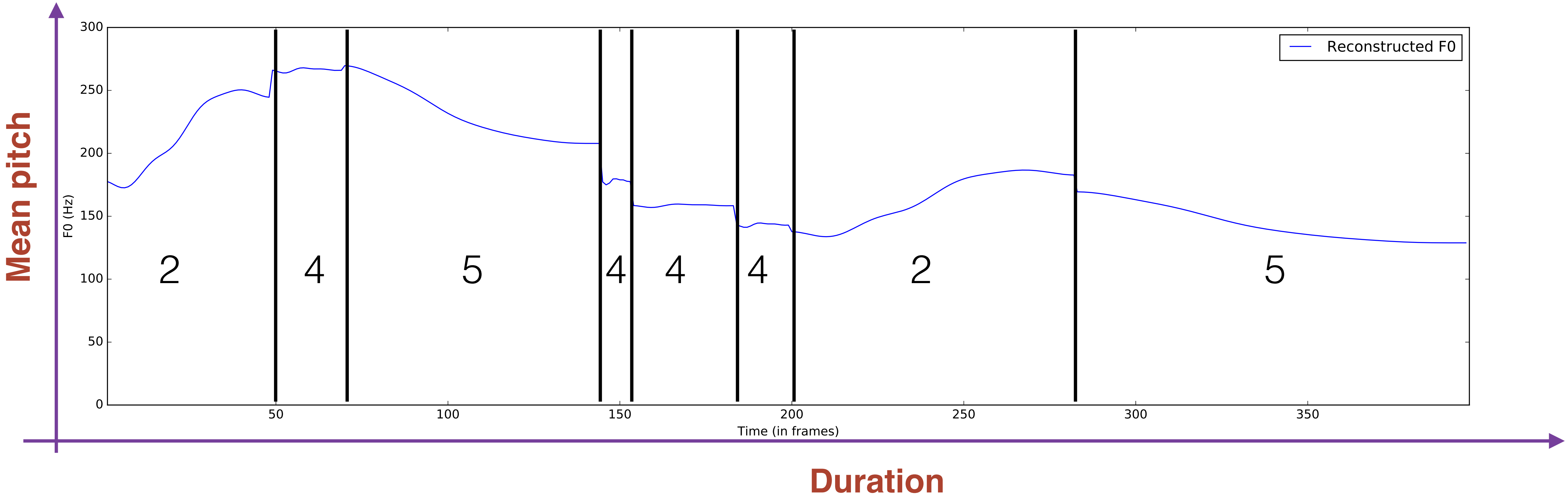


Intonation reconstruction from templates

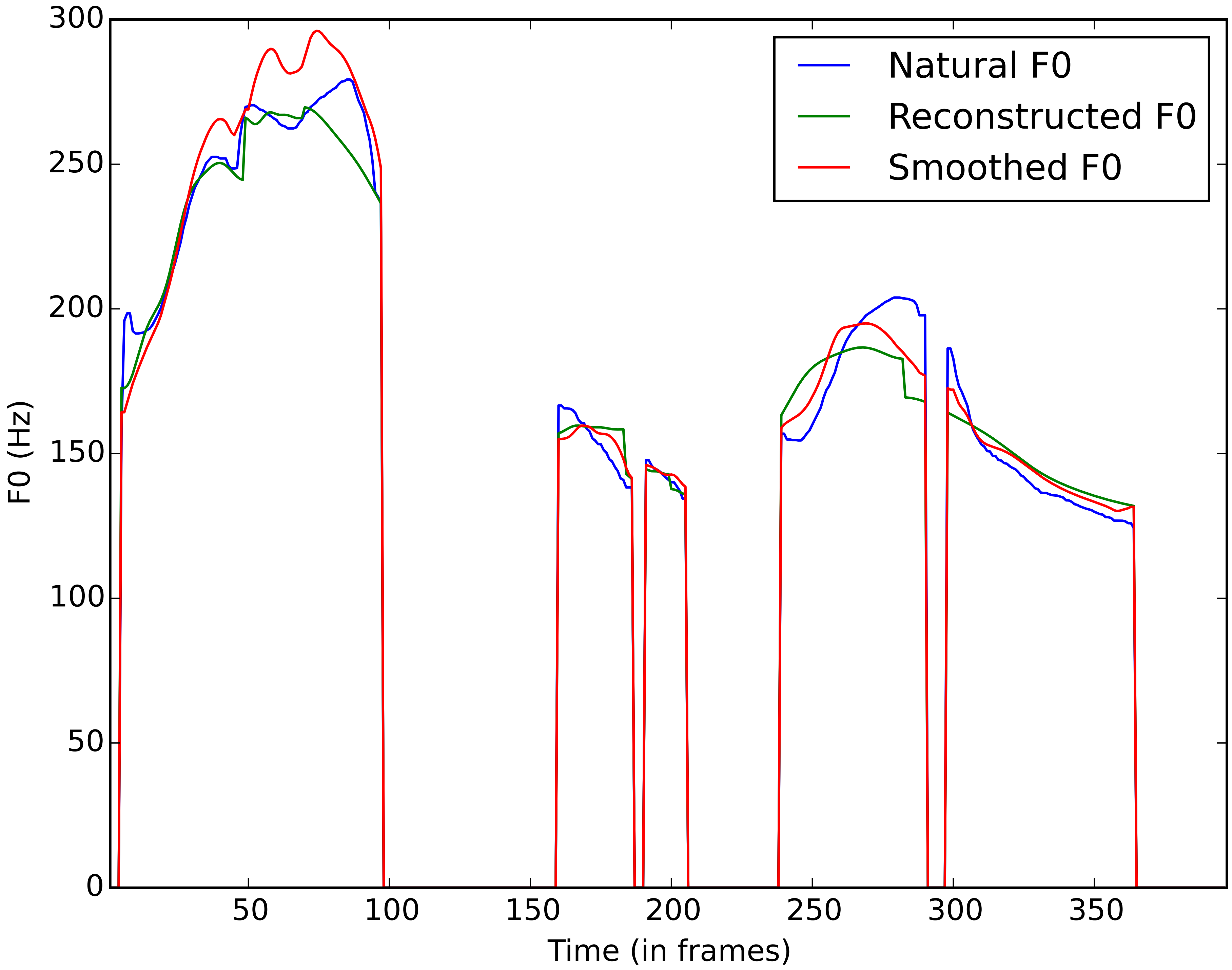
Goldilocks and the three bears

/g@Ulwd/	/lI/	/Qks/	/pau/	/and/	/D@/	/Tri/	/bE@z/
2	4	5	4	4	4	2	5

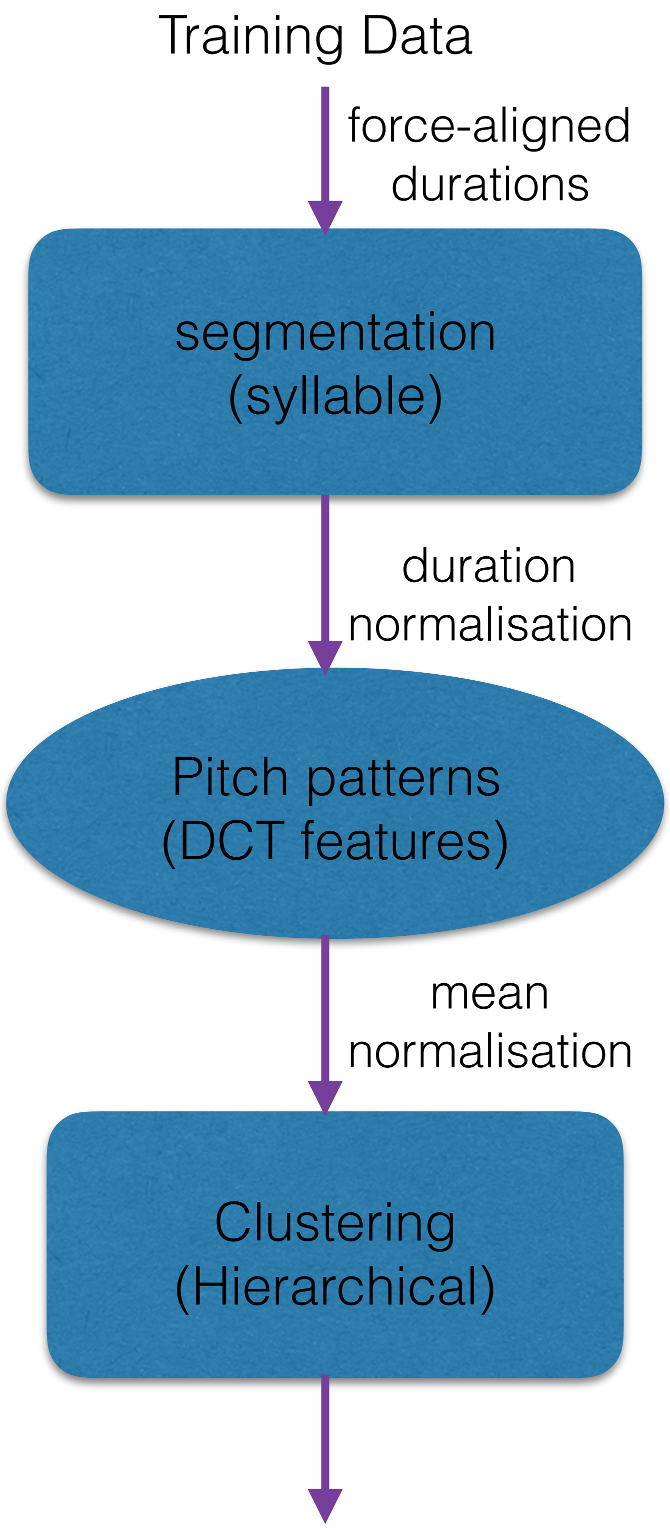




Intonation reconstruction



Hierarchical clustering - Recap

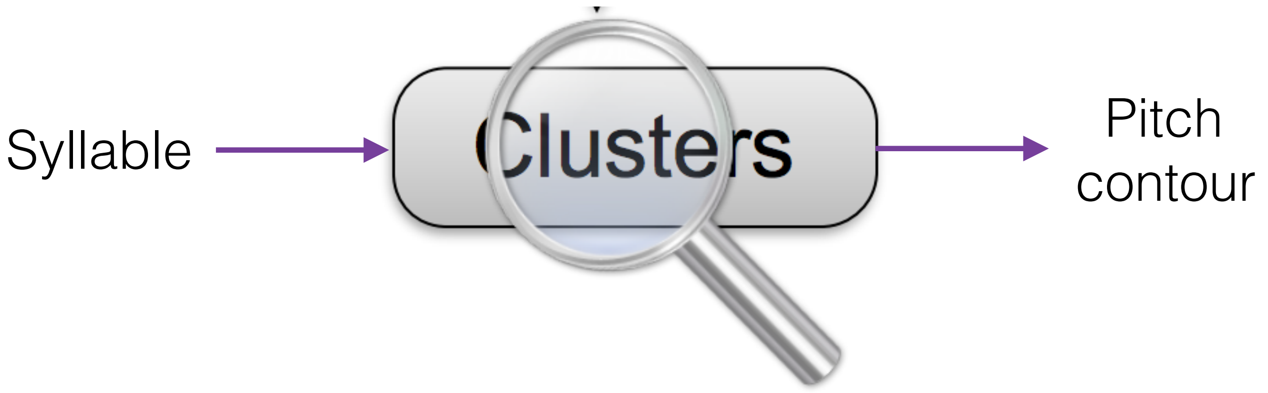


- Interpolate the F0 contour of each utterance and segment into syllables

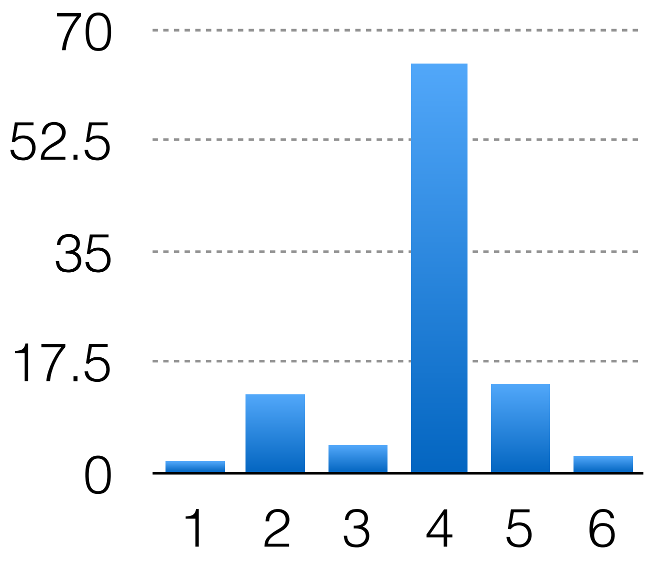
- Apply DCT based decomposition: c0 representing the mean over syllable, **c** = [c1,…,CN-1], representing the shape of the contour

- Perform top-to-bottom hierarchical clustering over the patterns (**c**).

Proposed approaches



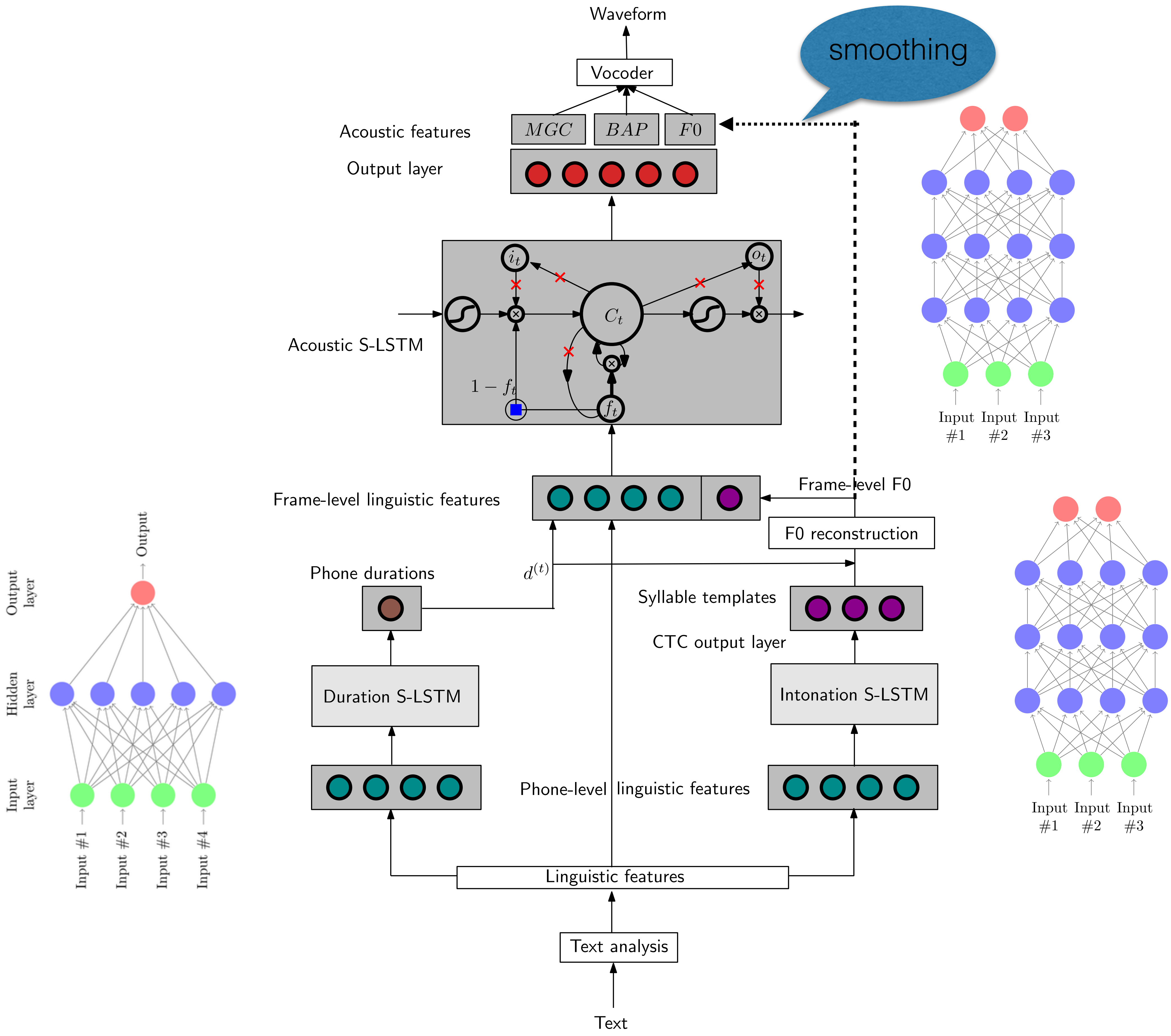
Neural Network classifiers:



template counts in the data

- A hierarchical deep neural network classifier (HC).
  - The first DNN choses between flat and non-flat template.
  - The second DNN choses among rest of the non-flat templates.

- A simplified LSTM with a CTC output layer (CTC).
  - Connectionist temporal classification coupled with S-LSTM to predict the sequence of templates given sequence of phonemes.


Waveform
Vocoder
Acoustic features
MGC
BAP
$F0$
Output layer
Acoustic S-LSTM
i_t
o_t
C_t
$1 - f_t$
f_t
Frame-level linguistic features
Frame-level F0
F0 reconstruction
Phone durations
$d^{(t)}$
Syllable templates
CTC output layer
Duration S-LSTM
Intonation S-LSTM
Phone-level linguistic features
Linguistic features
Text analysis
Text
Output
Output layer
Hidden layer
Input layer
Input #1
Input #2
Input #3
Input #4
Input #1
Input #2
Input #3
Input #1
Input #2
Input #3

smoothing

Results: systems

- Baseline system
  - MSE - A frame-wise regression baseline predicting F0 using LSTMs.
- Proposed systems
  - HC - A hierarchical deep neural network classifier
  - CTC - A simplified LSTM coupled with CTC output layer
  - Oracle - A oracle system using templates derived from natural F0 contour but with *predicted F0 mean and duration*

Objective evaluation

- Classification measures
  - Accuracy - percentage of templates correctly classified
  - F1 score - is a measure of test’s accuracy (precision and recall)

Model	Accuracy	F1 score
HC	61.1%	0.590
CTC	63.8%	0.593

Objective evaluation

- F0 prediction measures
  - RMSE - Root mean square error
  - CORR - Pearson correlation

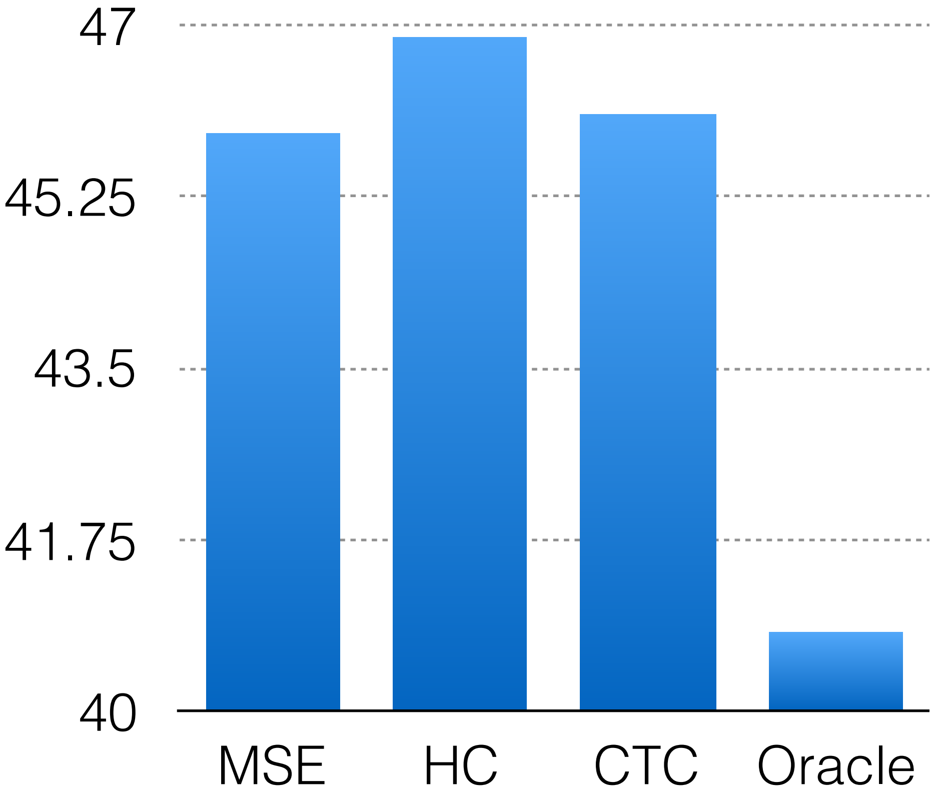


Fig: RMSE of predicted F0

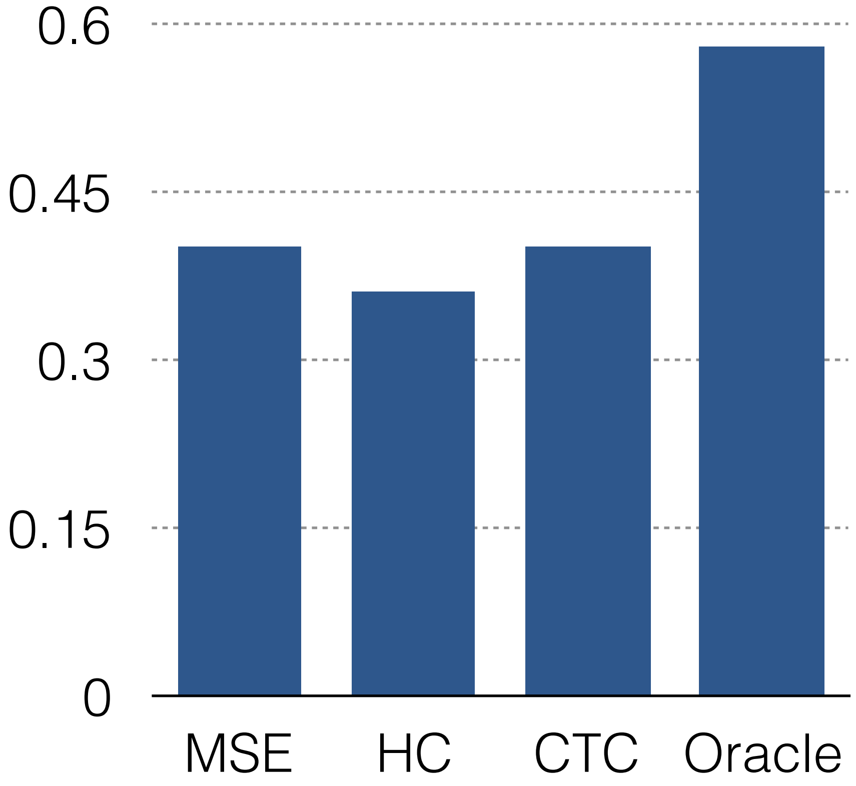


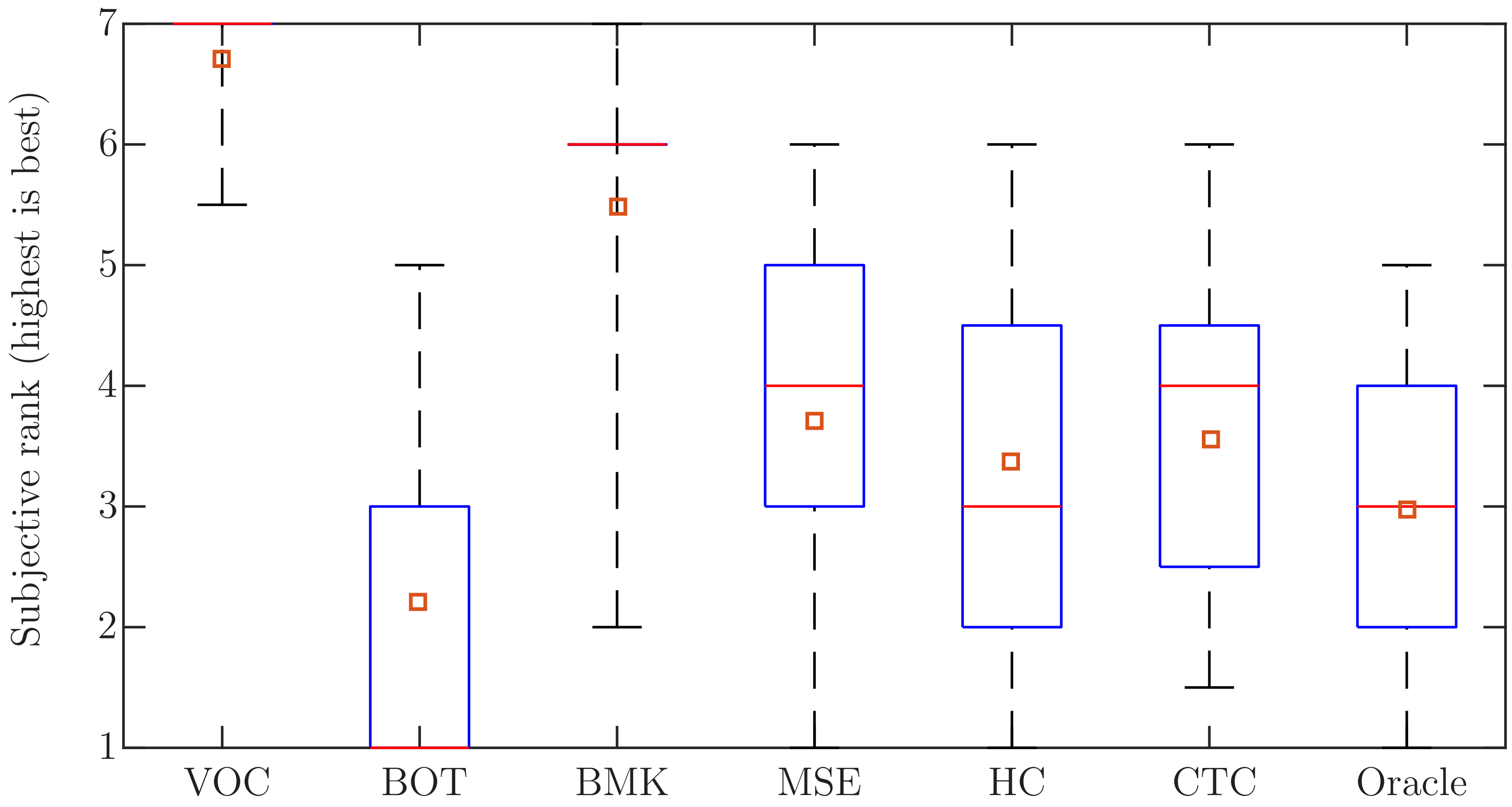
Fig: Correlation of predicted F0

- Oracle templates + Oracle F0 mean - 0.89 (corr.)

Subjective evaluation

- Reference systems
  - MSE - A frame-wise regression baseline predicting F0 using LSTMs.
  - BOT - A bottom line using piecewise-constant F0 per syllable (the mean natural F0)
  - BMK - A benchmark system using force-aligned durations and natural F0 contours
  - VOC - A top line of vocoded speech (STRAIGHT in this work)

Subjective evaluation: MUSHRA



- 20 listeners
- 20 out of 32 test stimuli

Fig: Box plot of aggregate ranks from listening test. Red lines are medians, orange squares means.

Summary and conclusions

- A classification approach to intonation prediction with syllable F0 templates
- Proposed approach matches the performance of conventional approach
- Has potential to exceed it once the issues with oracle template system are overcome
- Future work:
  - Better smoothing techniques and word-level templates
  - Use the prediction probabilities as features for frame-level regression approaches

Code

- Code for templates and clustering
  - https://github.com/ronanki/Hybrid_prosody_model
- Code for training neural networks
  - https://github.com/CSTR-Edinburgh/merlin