



Continuous Fundamental Frequency Prediction with Deep Neural Networks

Bálint Pál Tóth, Tamás Gábor Csapó
csapot@tmit.bme.hu

Budapest University of Technology and Economics

Introduction

Deep Learning: New era of machine learning

Feed forward deep neural networks

Speech research

- Speech recognition
- Speech coding
- Speech synthesis: using parametric vocoders
 - spectral components,
 - phone durations,
 - fundamental frequency (= pitch = F0).

Fundamental frequency prediction

Rule based F0 prediction

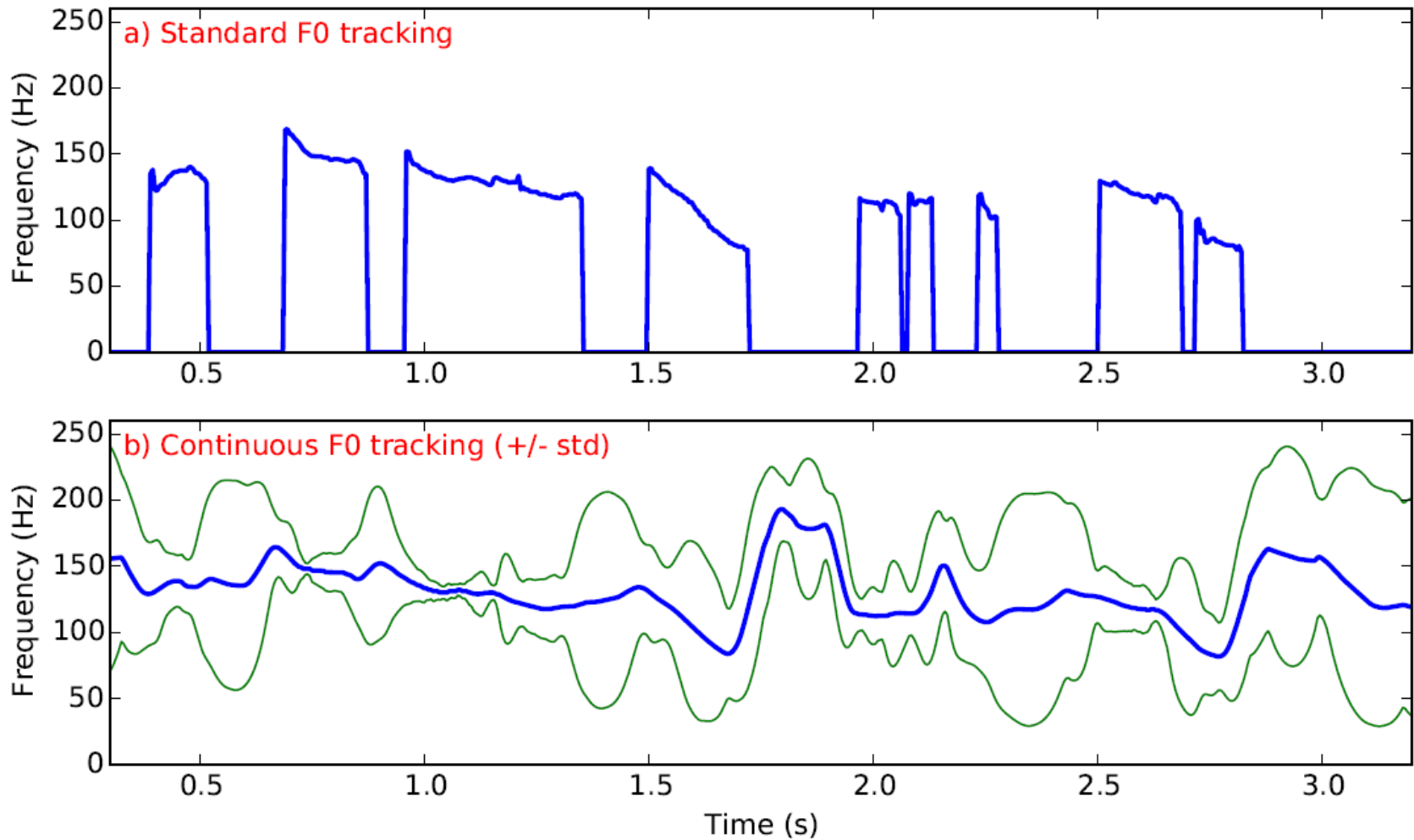
Statistical / machine learning approach

- Hidden Markov Models (HMM)
- Feed forward deep neural networks (DNN)

Pitch tracking algorithm

- Vanilla: Standard F0 tracking + voiced/unvoiced tagging
 - Difficulty in modeling standard F0
 - Discontinuity in unvoiced regions
 - Multi-Space Distribution Hidden Markov Models (MSD-HMM)
- Proposed: Continuous F0 + Maximum Voiced Frequency
 - No discontinuity, less difficulty in modeling

Continuous Pitch Tracking



'I saw it all myself, and it was splendid.'

Goal

Investigation of

- 1) feed forward deep neural networks modeling power,
- 2) model complexity of vanilla and continuous F0 trajectories

Hypothesis

Perceptual quality of DNN-based prediction using continuous F0 will be superior to discontinuous F0

Vocoder methods I: Standard F0 (baseline)

Pulse-noise vocoder

SWIPE pitch tracking algorithm (Camacho & Harris 2008)

2 parameters for every 25 ms long (5 ms shift) window:

- F0 value for voiced regions
 - For DNN, linear interpolation in unvoiced regions
- Voiced / unvoiced binary flag

Denoted by F0std

Vocoder methods II: Continuous F0

Residual-based continuous vocoder

SSP pitch tracking algorithm (Garner et al., 2013)

2 parameters for every 25 ms long (5 ms shift) window:

- F0 value for all regions
- Maximum Voiced Frequency (MVF)
 - Voiced-unvoiced frequency boundary

Denoted by F0cont

Machine Learning Methods: HMM

Widespread statistical parametric speech synthesis approach

Vocoder I

- F0std training and prediction (with MSD-HMM)

Vocoder II

- F0cont & MVF training and prediction

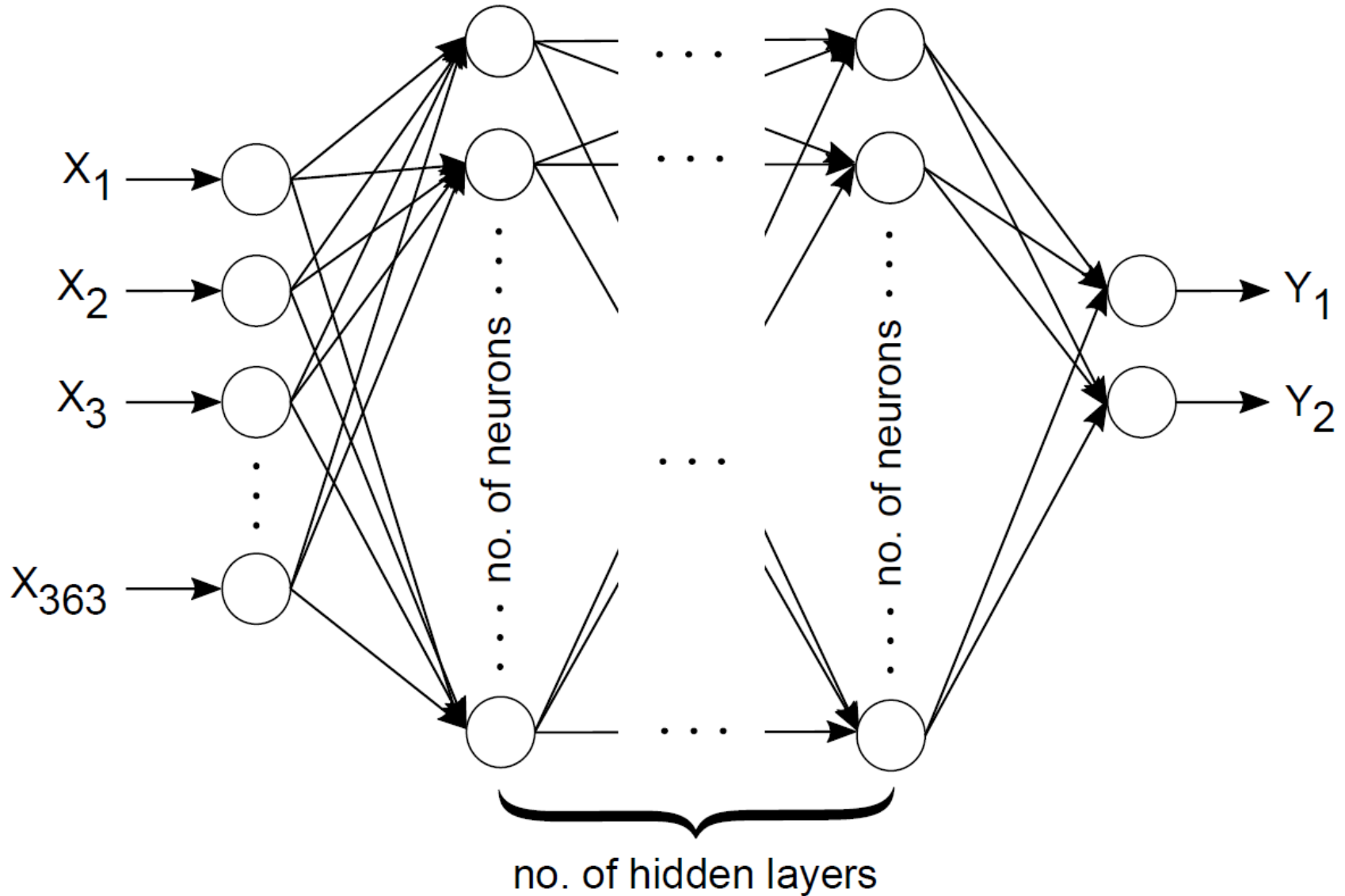
HTS 2.3 with default settings

Machine Learning Methods: DNN

Feed forward deep neural networks

- Mean Square Error (MSE) cost function
- ADADELTA optimization with mini-batches
- Parametric Rectified Linear Units (PReLU) as activation function for hidden layers
- Sigmoid activation function for the outputs
- Weight initialization:
 - Xavier's weight input-hidden and hidden-output layers
 - Orthogonal in the hidden layers
- Dropout w/ 50% after each layer except output layer
- Early stopping after 50 epochs

Proposed DNN network



DNN Inputs

Parameter-wise transformed
to zero mean and unit variance

Feature name	#	Type
Quinphone	5*68	One-hot
Number of phonemes/syllables/words/phrases in the previous/current/next syllable/word/phrase/sentence	4*3	Numerical
Number of syllables/words in the current sentence	2	Numerical
Forward/backward position of the actual phoneme/syllable/word/phrase in the syllable/word/phrase/sentence	2*3	Numerical
Phone boundaries	2	Numerical
Percentual position of the actual frame within the phone	1	Numerical
Altogether:	363	

DNN Outputs

Normalized to 0.01...0.99 for sigmoid activation

System	Feature name	#	Type
F0std	LogF0	1	Continuous (interpolated)
	V/UV flag	1	Binary
F0cont	LogF0	1	Continuous
	MVF	1	Continuous

Evaluation: hyperparameter optimization

One male and one female speaker from
Precisely Labelled Hungarian Database (PLHD)

1984 utterances / speaker (~2 hours)

Training-validation-test sets: 80-15-5%

Hyperparameter optimization with male speaker:

- #hidden layers: 1..7
- #neurons / layer: 80..2048
- #mini-batch size: 8..256

Validation loss was measured.

64 neural nets for F0std and 73 for F0cont

Top 5 were selected and run with female speaker

Hyperopt results: standard F0

ID	# Hidden Layers	# Neurons	Epochs	Validation MSE
F0std-1	3	350	61	0.01076
F0std-2	3	650	32	0.01078
F0std-3	3	900	30	0.01089
F0std-4	3	950	36	0.01099
F0std-5	3	800	37	0.01103

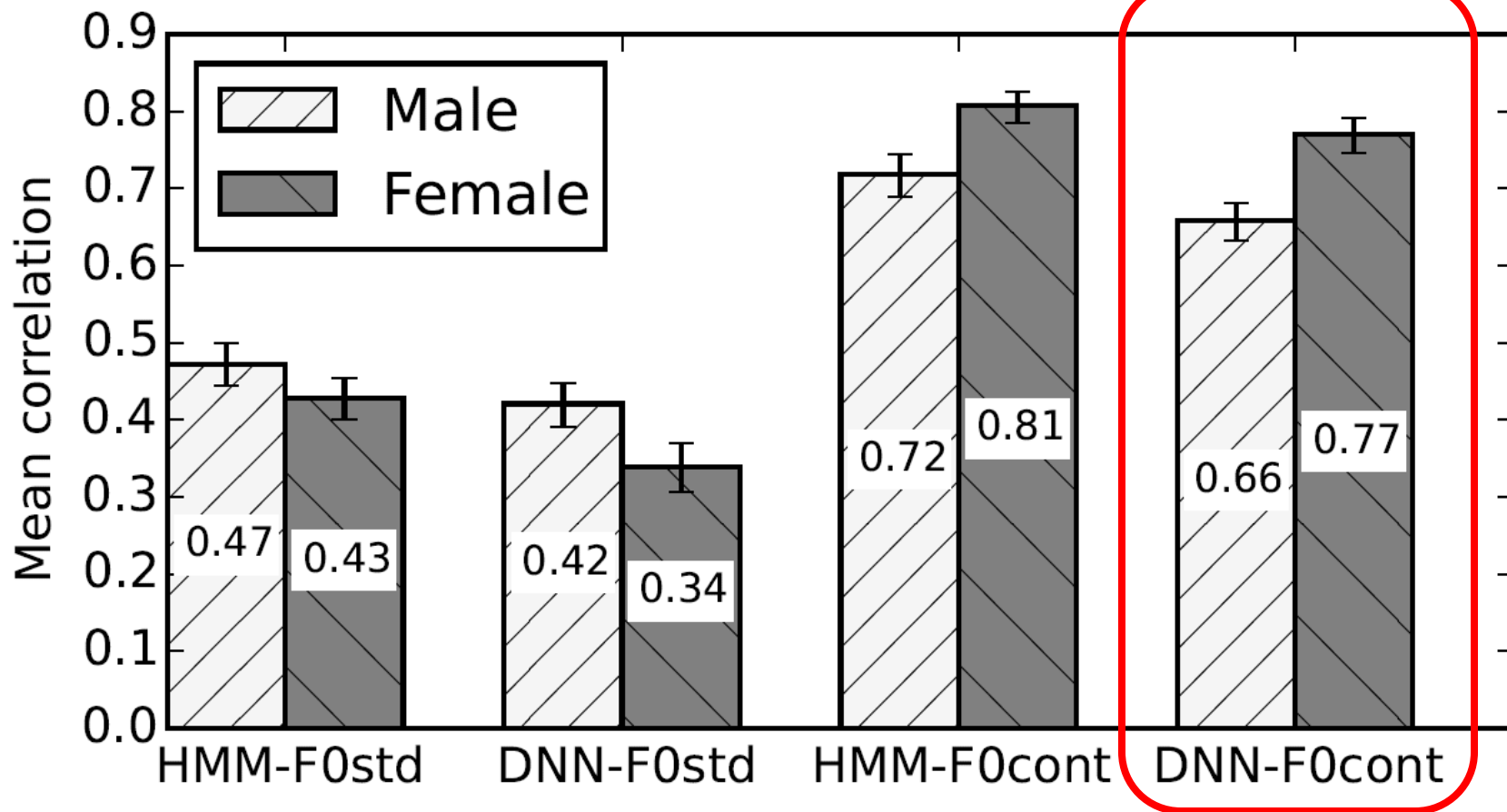
Mini-batch size = 128

Hyperopt results: continuous F0

ID	# Hidden Layers	# Neurons	Epochs	Validation MSE
F0cont-1	3	160	2	0.00239
F0cont-2	3	80	67	0.00346
F0cont-3	1	128	2	0.00349
F0cont-4	3	70	12	0.00352
F0cont-5	2	100	28	0.00356

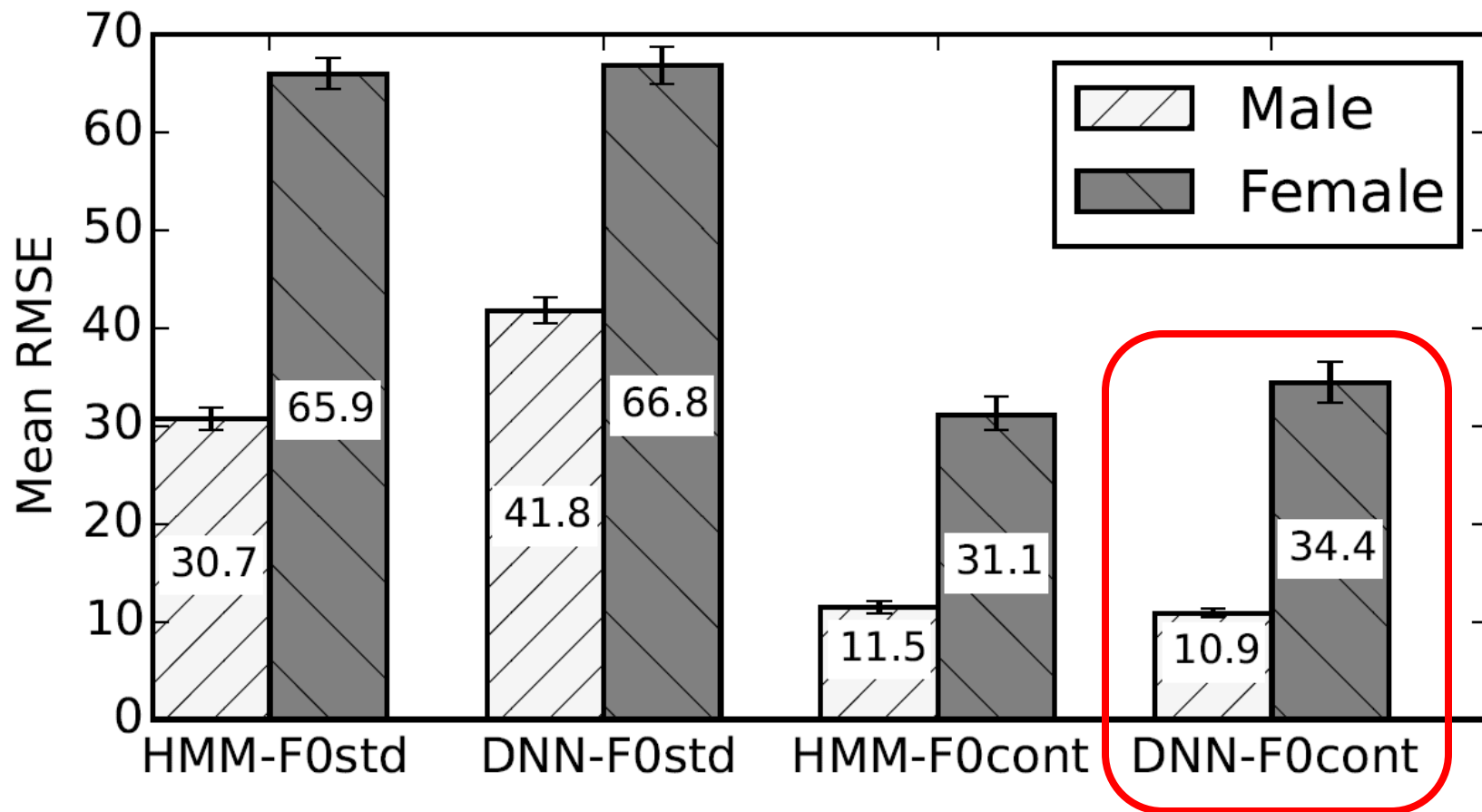
Mini-batch size = 8

Objective evaluation I



Mean correlation between natural F0 and modeled F0
(higher value: larger similarity between
compared F0 trajectories)

Objective evaluation II



Mean RMSE between natural F0 and modeled F0
(higher value: larger difference between
compared F0 trajectories)

Subjective evaluation I

Goal: measure the perceived intonation of sentences

Web-based MUSHRA test:

- Reference natural sentence,
- Vcoded sentence with F0 from
 - Natural utterance
 - F0std
 - F0cont
 - HMM
 - F0std
 - F0cont
 - DNN
 - F0std
 - F0cont
- Benchmark: vocoded with F0=0

Subjective evaluation II

Sentences with highest RMSE were selected

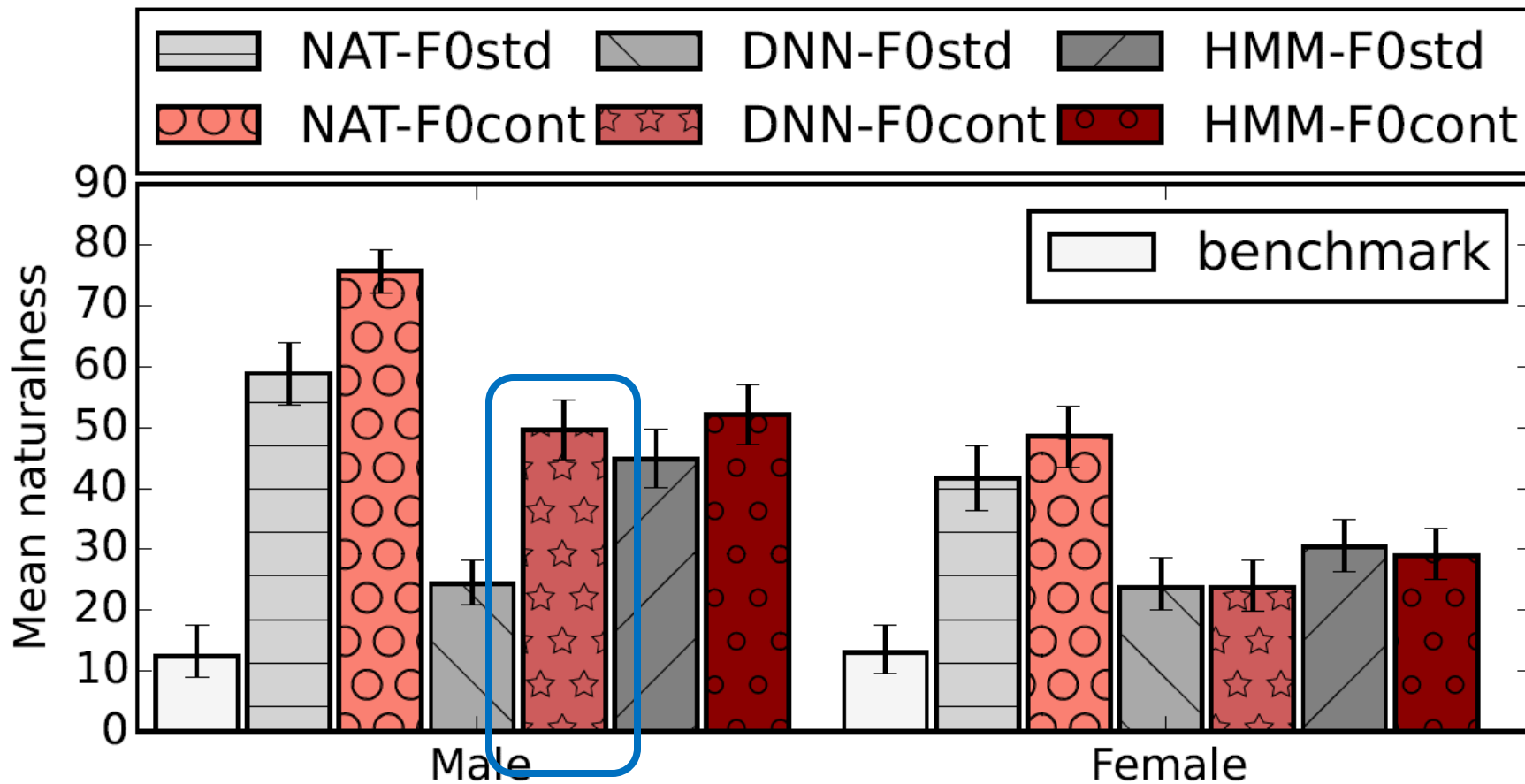
2 speakers × 8 types × 5 sentences (altogether 80 sent.)

Randomized order

18 test subjects (9 females, 9 males)

13 minutes to complete the test (avg)

Subjective evaluation III



(higher value: more similar to natural)

Conclusions and discussion

- 1) F0cont can be predicted better than F0std with HMMs and DNNs
- 2) Simpler DNN models for F0cont (good for embedded systems)
- 3) F0cont has faster convergence (we measured cca. 7x faster than F0std)
- 4) Simple DNN approaches the F0 modeling capacity of state-of-the-art HMM

→ continuous representation of F0 forms a less complex system than the V/UV based F0std



M Ű E G Y E T E M 1 7 8 2

Thanks for listening!

`csapot@tmit.bme.hu`