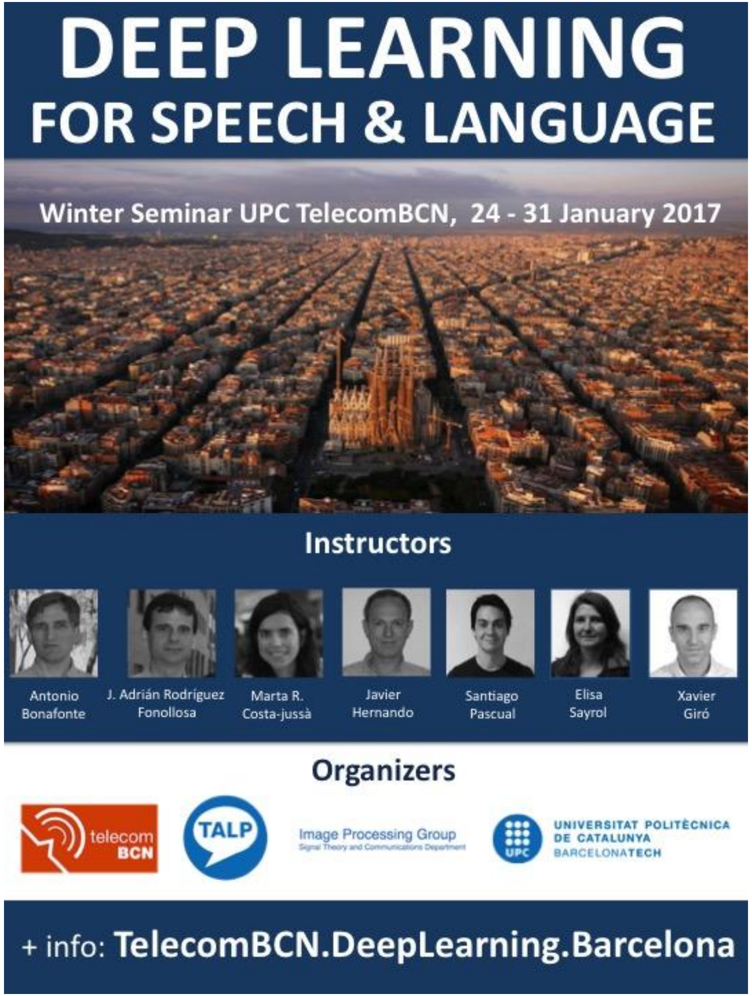


[course site]

Day 3 Lecture 5

Parametric Speech Synthesis

Antonio Bonafonte

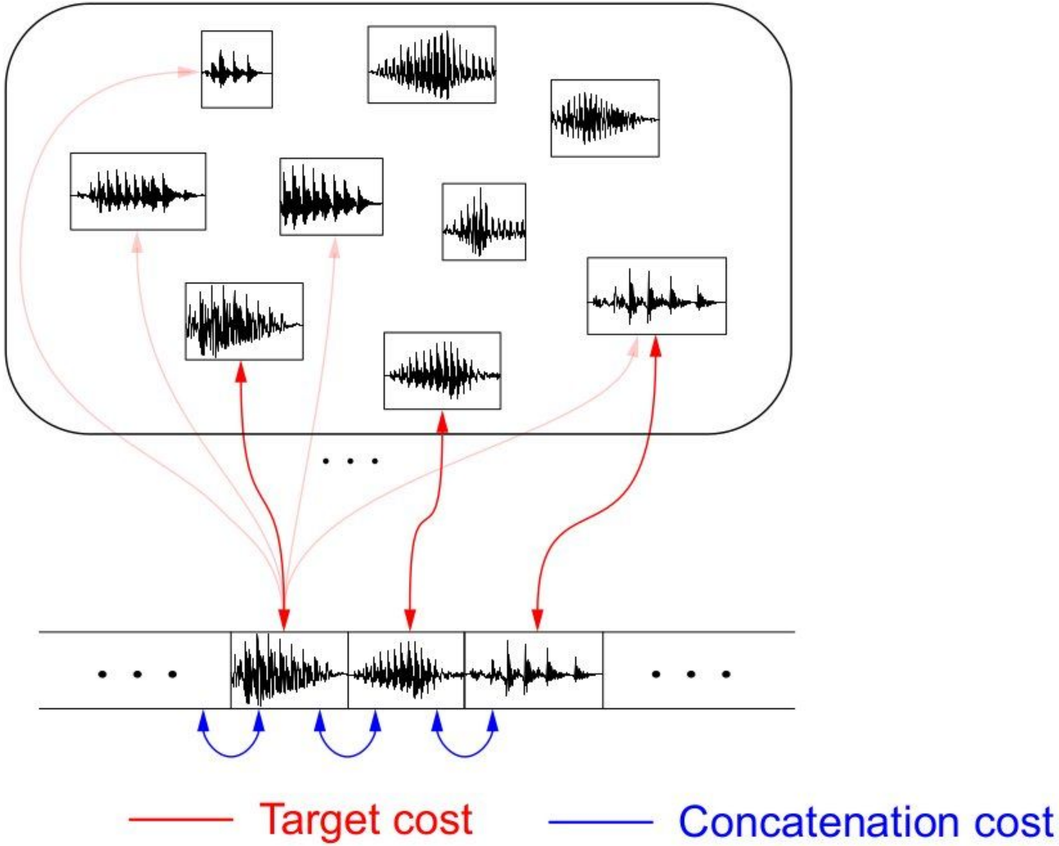
Main TTS Technologies

Concatenative speech synthesis + Unit Selection

Concatenate *best* prerecorded speech *units*

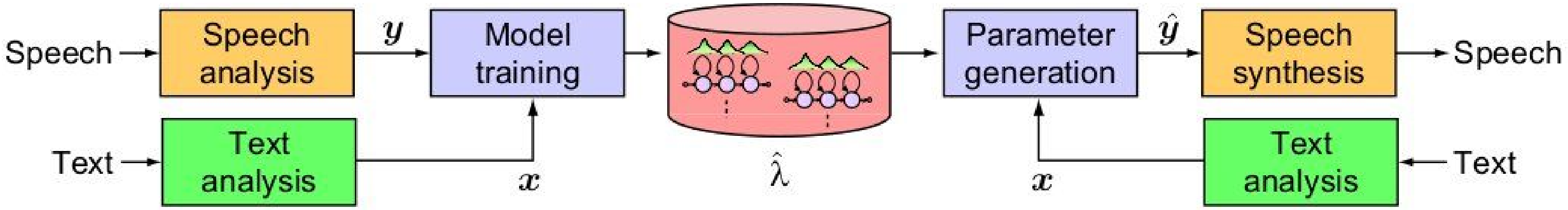
Speech data: 2-10 hours, professional speaker, carefully segmenten and annotated.

Concatenative



H. Zen - RTTHSS 2015
http://rtthss2015.talp.cat/

Statistical Speech Synthesis



H. Zen - RTTHSS 2015
http://rtthss2015.talp.cat/

Main TTS Technologies

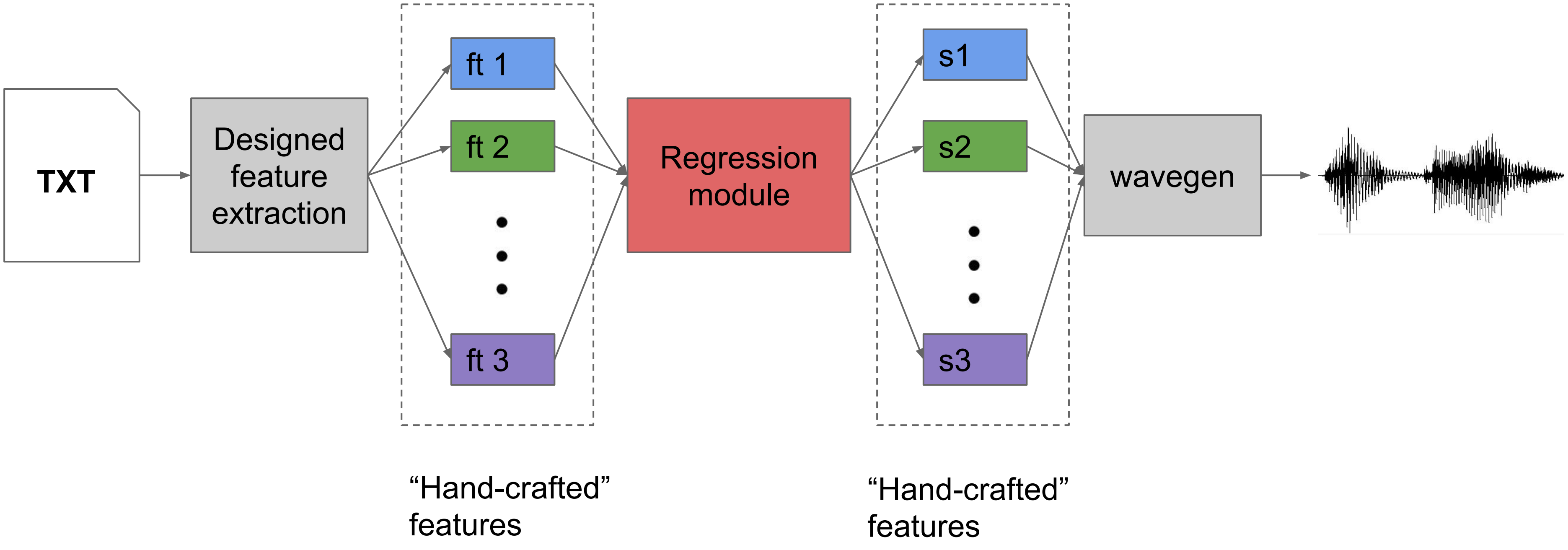
Concatenative speech synthesis + Unit Selection
- Concatenate *best* pre-recorded speech *units*

Statistical Parametric Speech Synthesis
- Represent speech waveform using parameters (eg 5ms)
- Use statistic generative model
- Reconstruct waveform from generated parameters

Hybrid Systems
- Concatenative speech synthesis
- Select *best* units attending a statistical parametric system

Deep architectures ... but not deep (yet)

Text to Speech: Textual features → Spectrum of speech (many coefficients)



Textual features (x)

From text to phoneme (pronunciation)

- Disambiguation, pronuntiation (e.g.: Jan. 26)

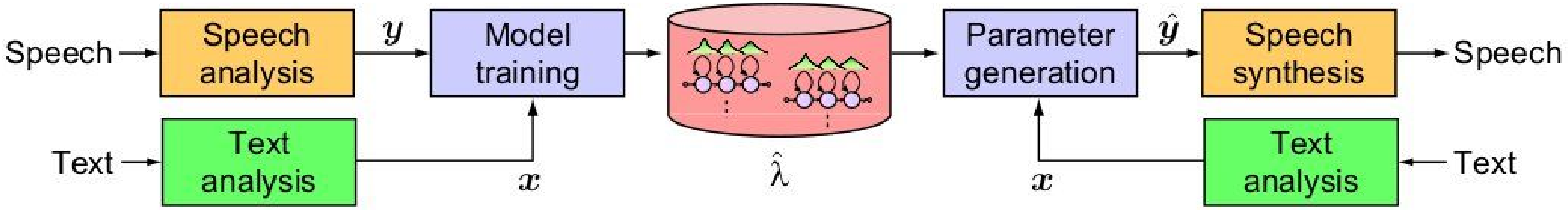
From phoneme to phoneme+ (with linguistic features)

Textual features (x)

- {preceding, succeeding} two phonemes
- Position of current phoneme in current syllable
- # of phonemes at {preceding, current, succeeding} syllable
- {accent, stress} of {preceding, current, succeeding} syllable
- Position of current syllable in current word
- # of {preceding, succeeding} {stressed, accented} syllables in phrase
- # of syllables {from previous, to next} {stressed, accented} syllable
- Guess at part of speech of {preceding, current, succeeding} word
- # of syllables in {preceding, current, succeeding} word
- Position of current word in current phrase
- # of {preceding, succeeding} content words in current phrase
- # of words {from previous, to next} content word
- # of syllables in {preceding, current, succeeding} phrase

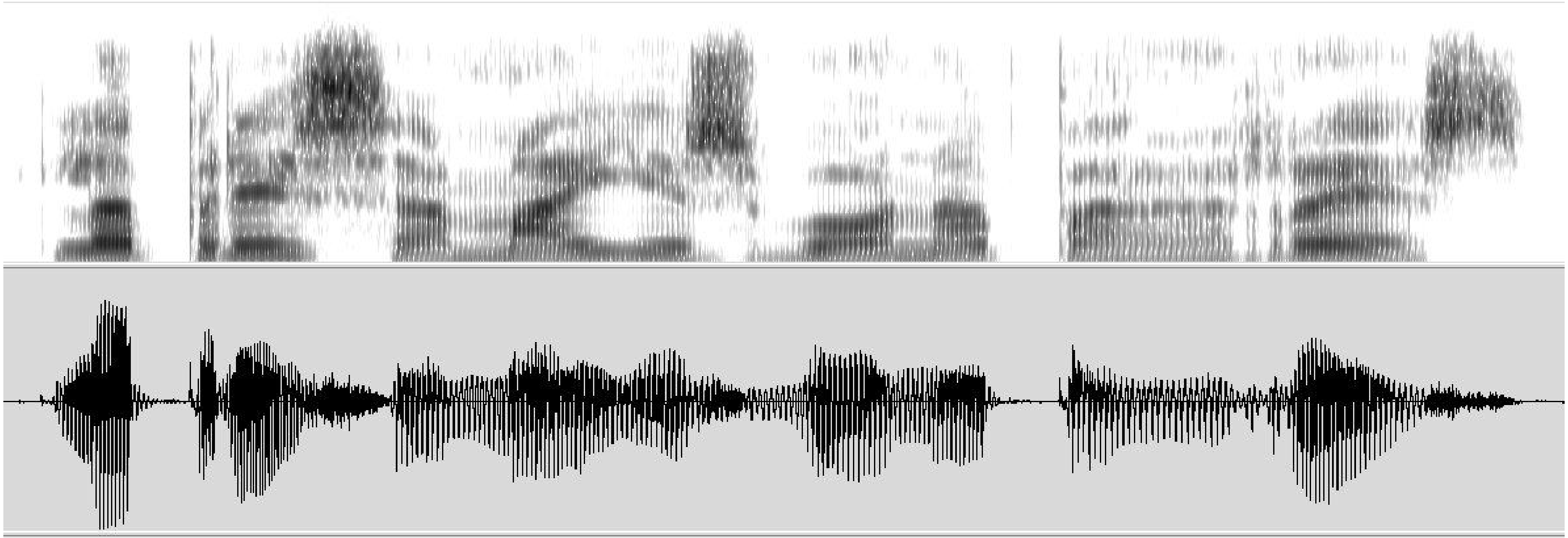
H. Zen - RTTHSS 2015
http://rtthss2015.talp.cat/

Statistical Speech Synthesis

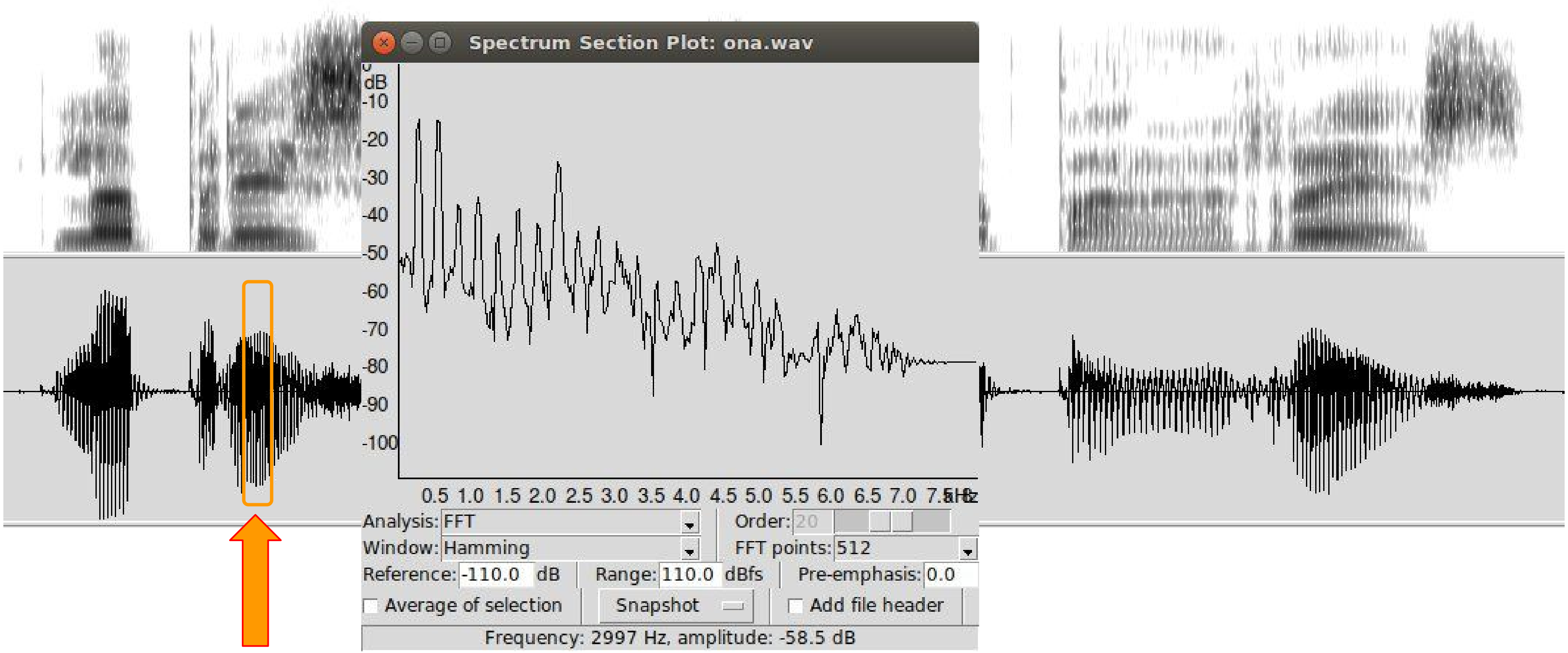


H. Zen - RTTHSS 2015
http://rtthss2015.talp.cat/

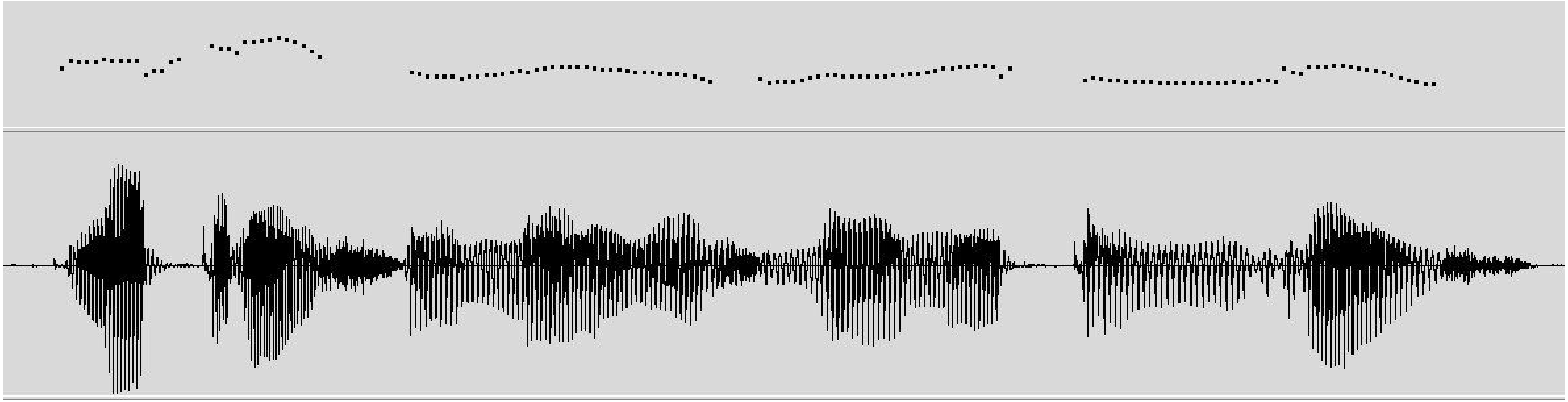
Speech features (*y*)



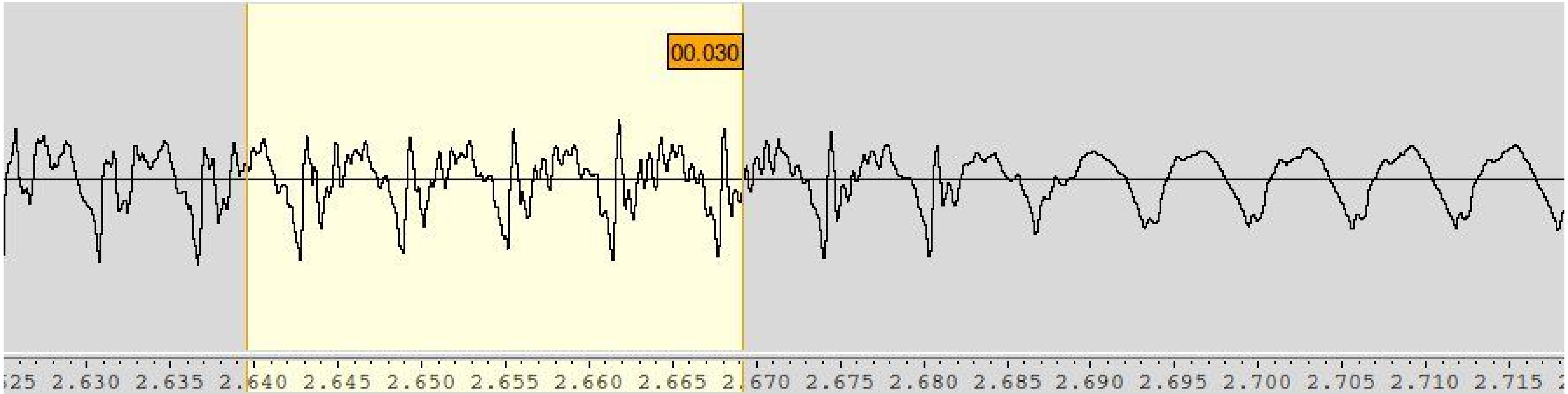
Speech features (y)



Speech features (*y*)

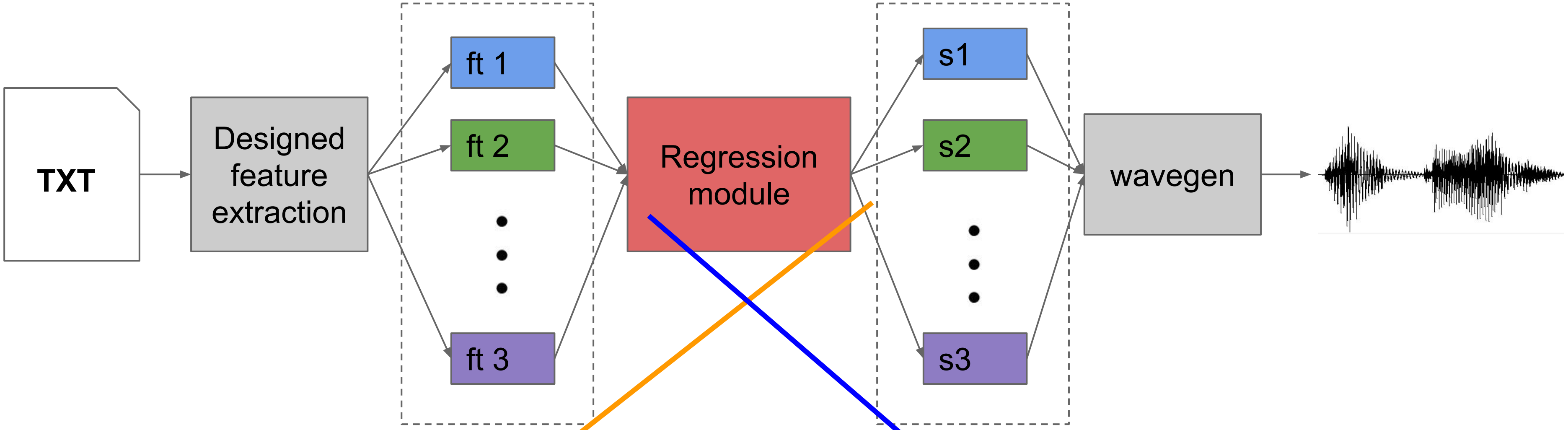


Speech features (*y*)



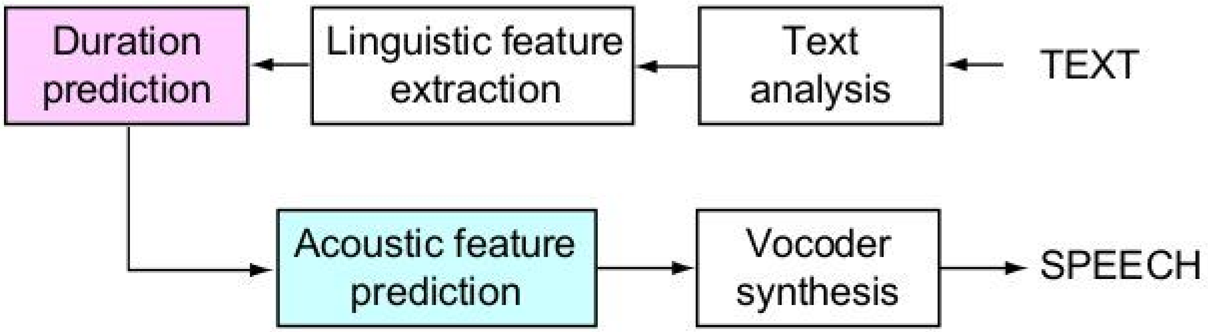
- Rate: ~ 5 ms. (200Hz)
- Spectral features (envelope)
- Excitation features (fundamental frequency, pitch)
- Representation that allows reconstruction: vocoders (Straight, Ahocoder, ...)

Regression

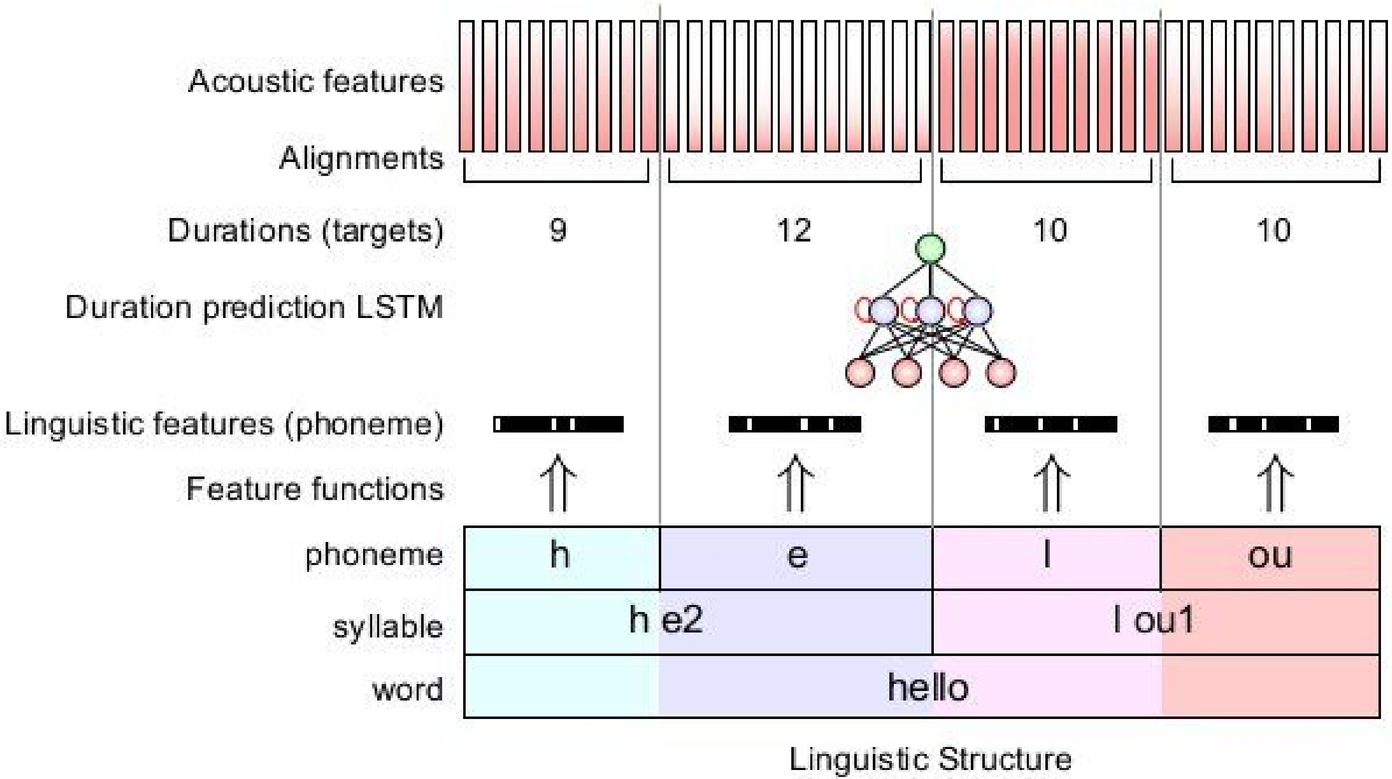


$$\hat{y} = \underset{y}{\operatorname{argmax}}\, p(y|x)$$

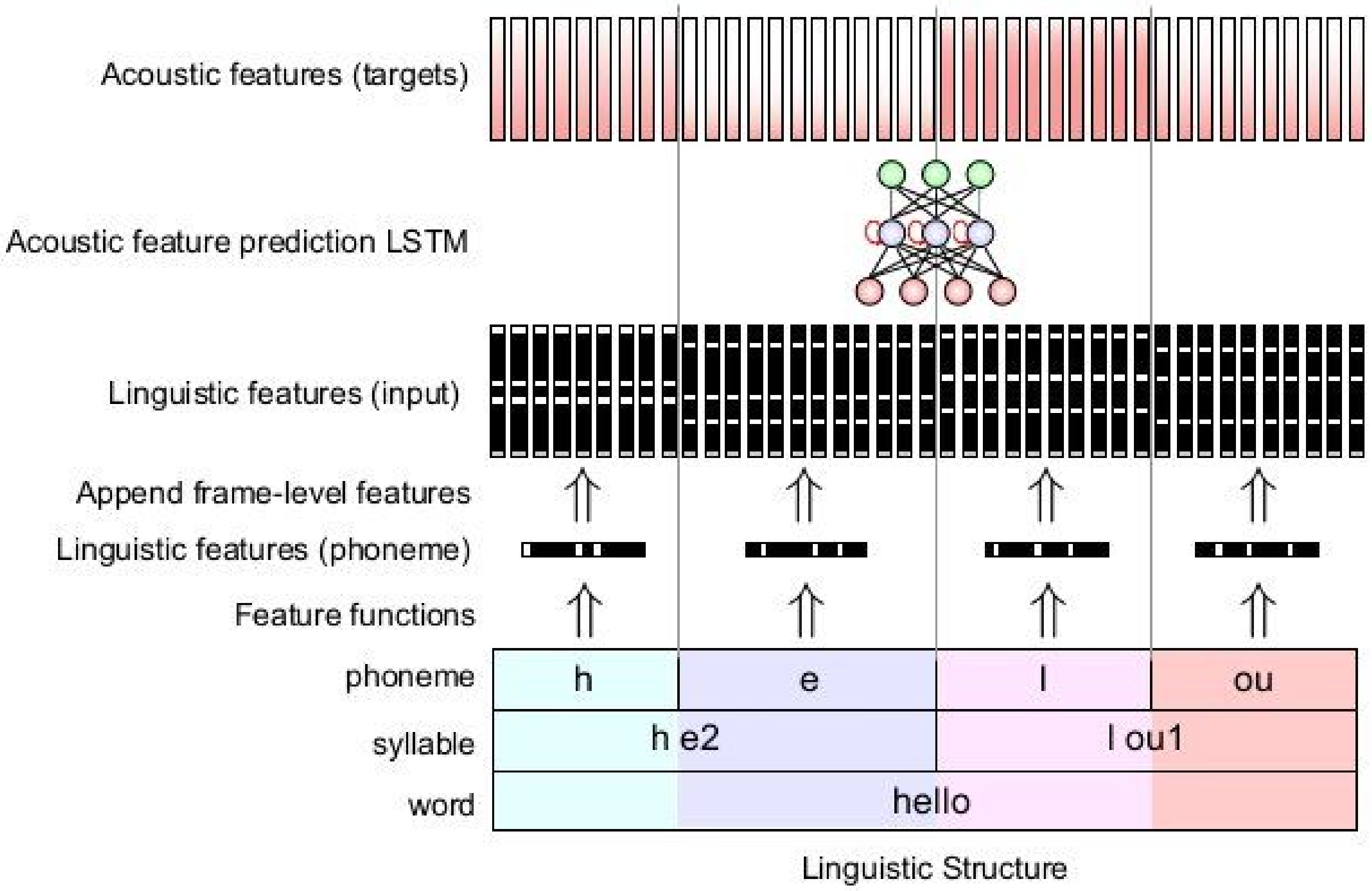
Phoneme rate vs. frame rate



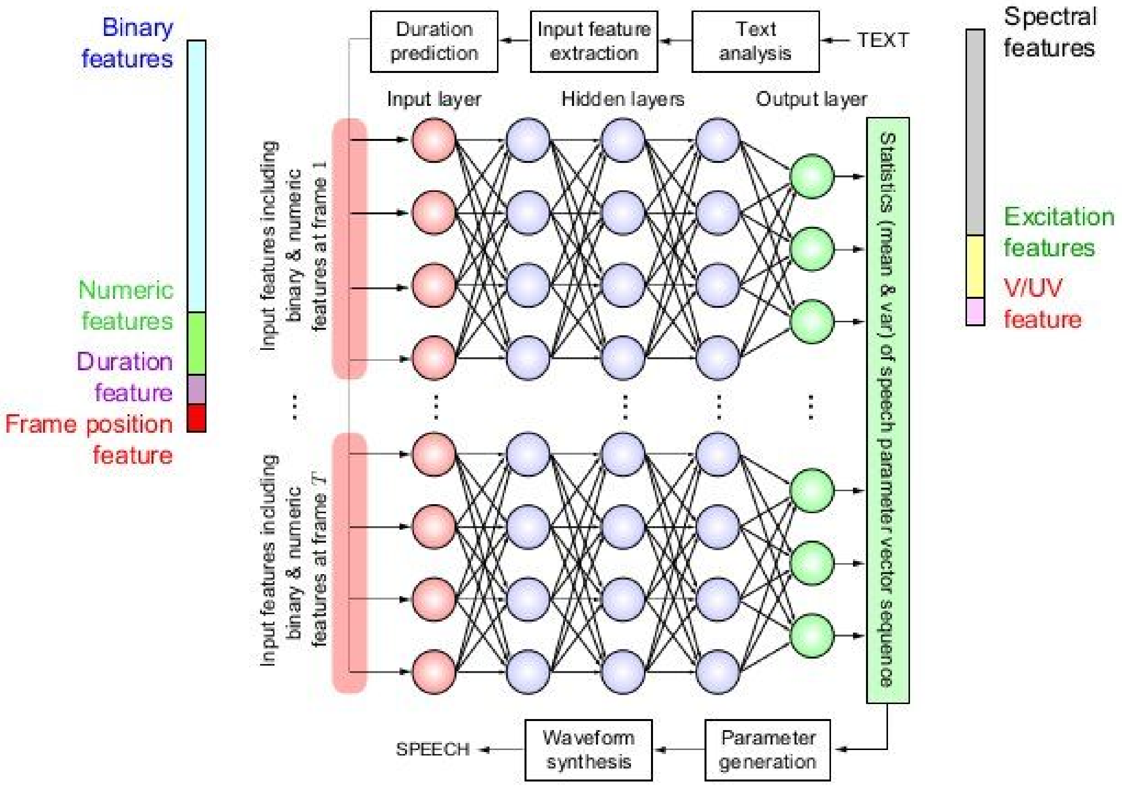
Duration Modeling



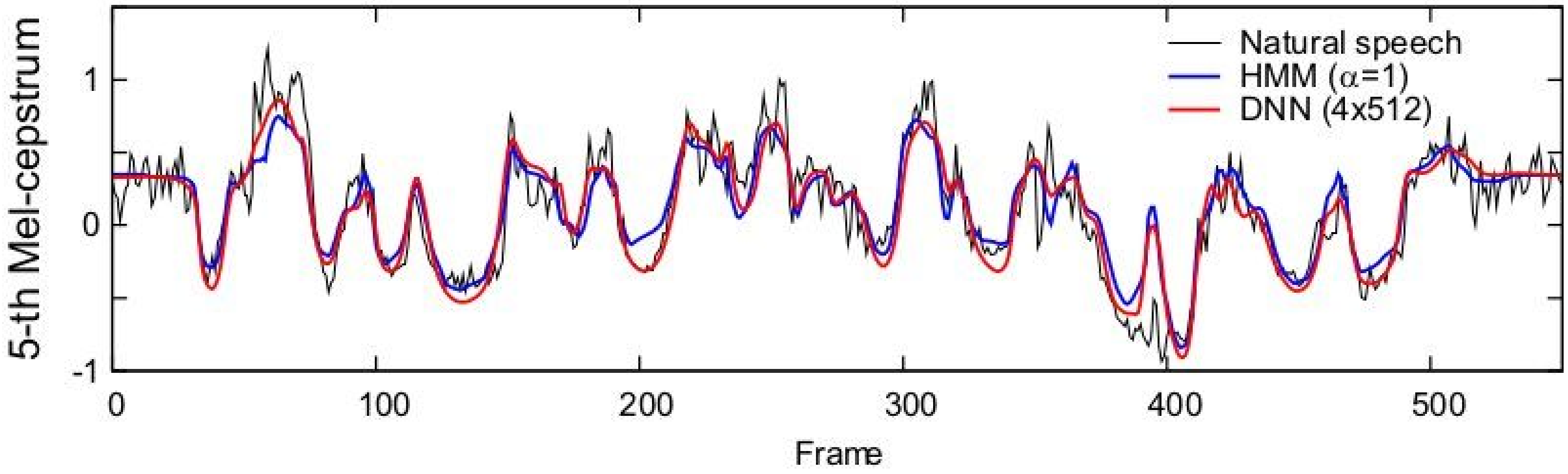
Acoustic Modeling



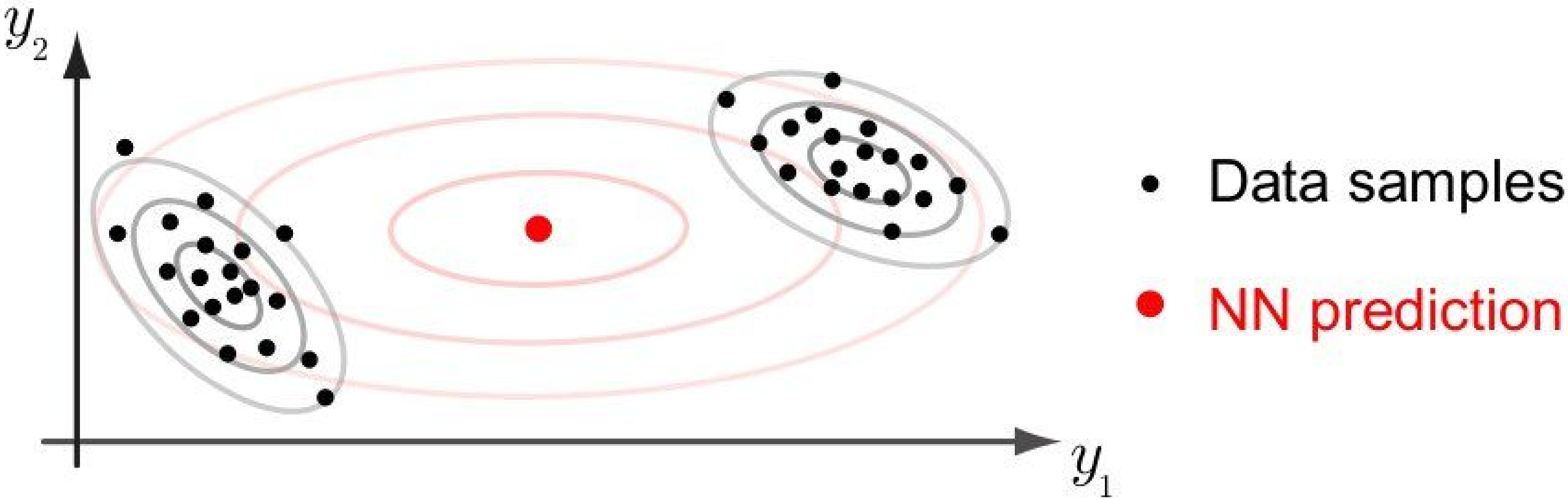
Acoustic Modeling: DNN



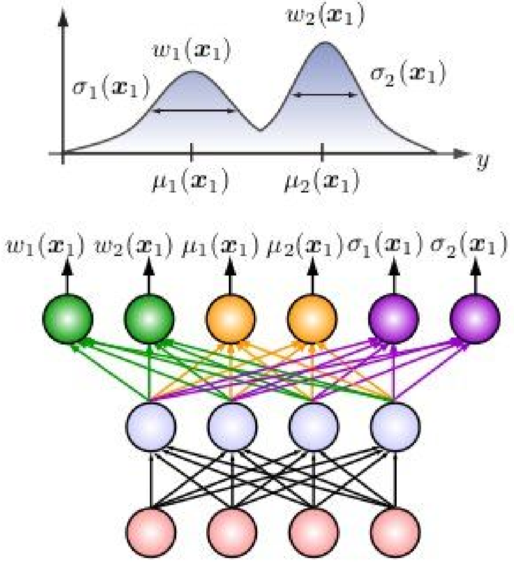
Acoustic Modeling: DNN



Regression using DNN (problem)



Mixture density network (MDN)



1-dim, 2-mix MDN

Inputs of activation function

$$z_j = \sum_{i=1}^{4} h_i w_{ij}$$

: Weights $\rightarrow$ Softmax activation function

$$w_1(\boldsymbol{x}) = \frac{\exp(z_1)}{\sum_{m=1}^{2} \exp(z_m)} \qquad w_2(\boldsymbol{x}) = \frac{\exp(z_2)}{\sum_{m=1}^{2} \exp(z_m)}$$

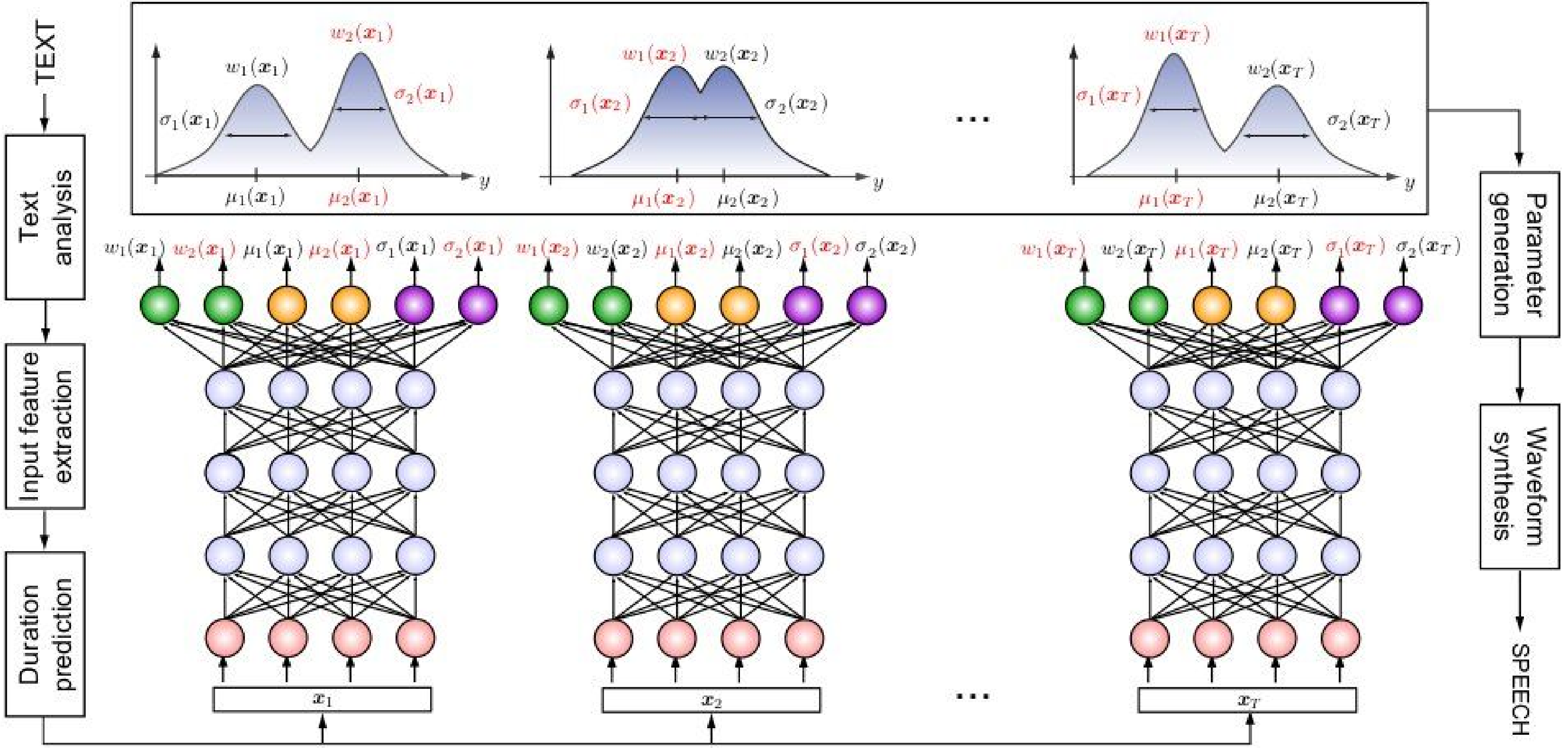
: Means $\rightarrow$ Linear activation function

$$\mu_1(\boldsymbol{x}) = z_3 \qquad \mu_1(\boldsymbol{x}) = z_4$$

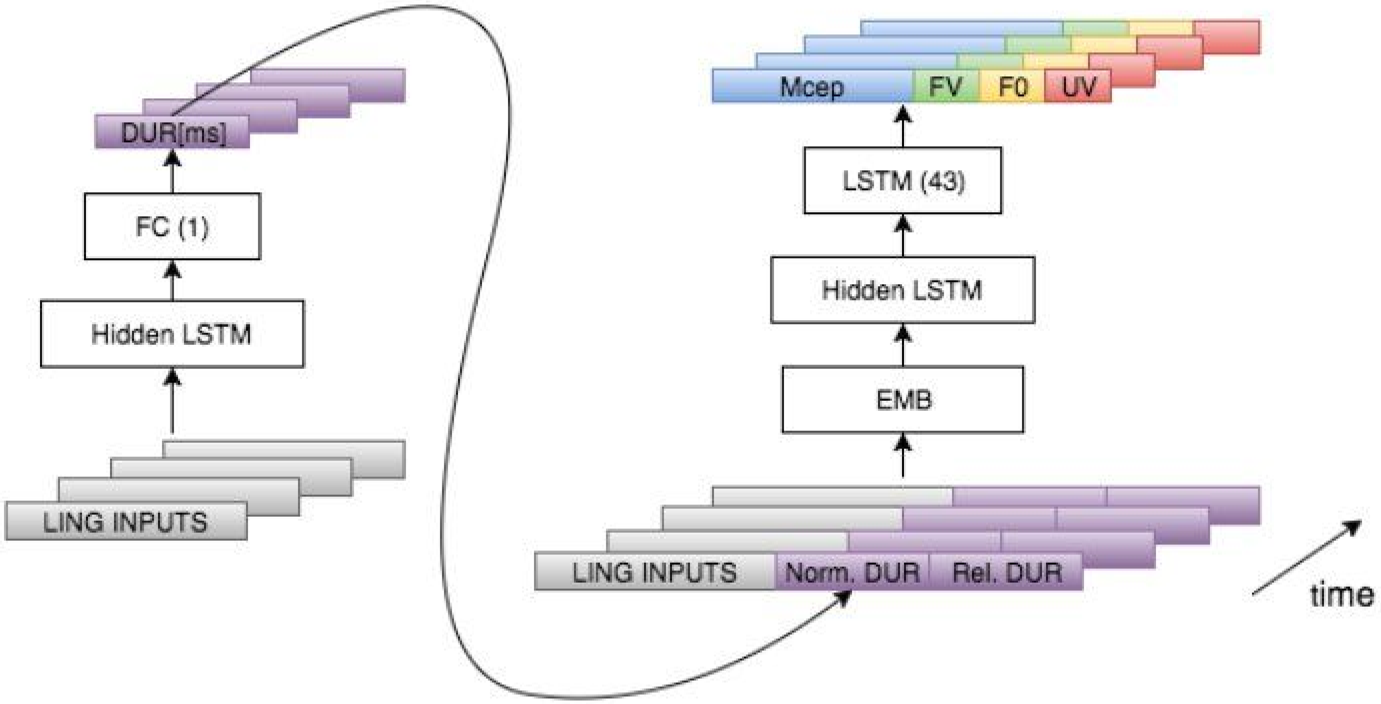
: Variances $\rightarrow$ Exponential activation function

$$\sigma_1(\boldsymbol{x}) = \exp(z_5) \qquad \sigma_2(\boldsymbol{x}) = \exp(z_6)$$

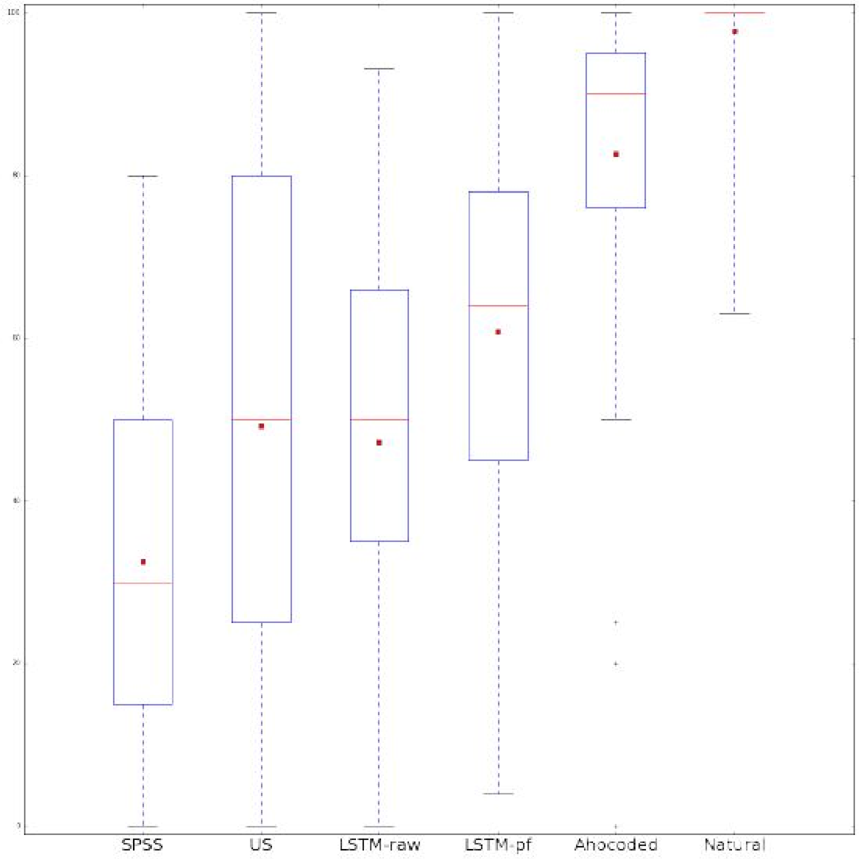
Mixture density network (MDN)



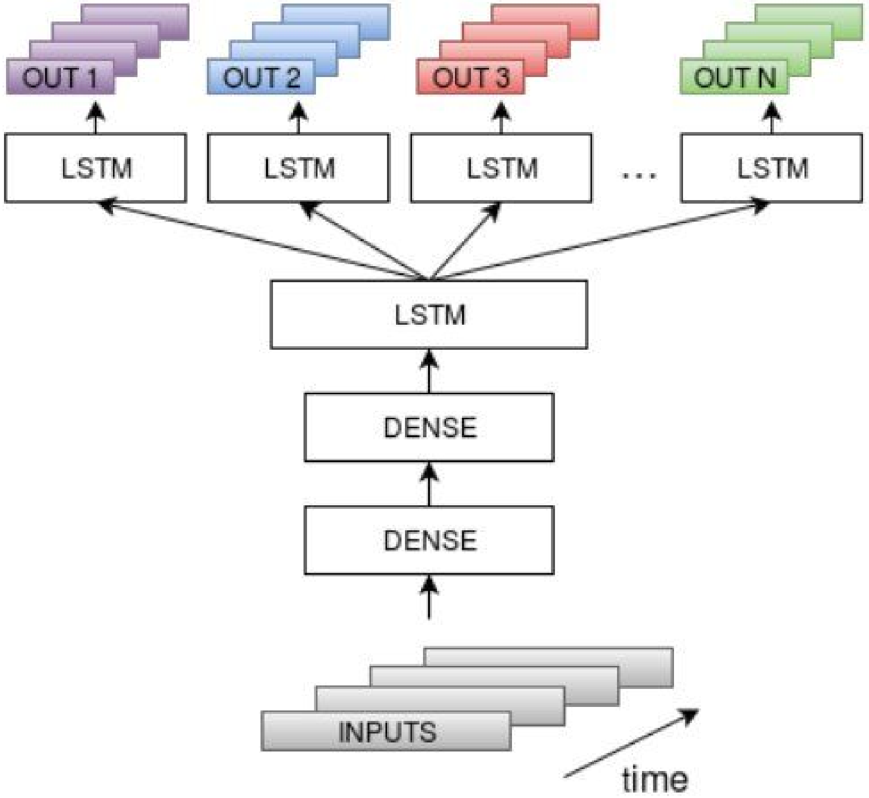
Recurrent Networks: LSTM



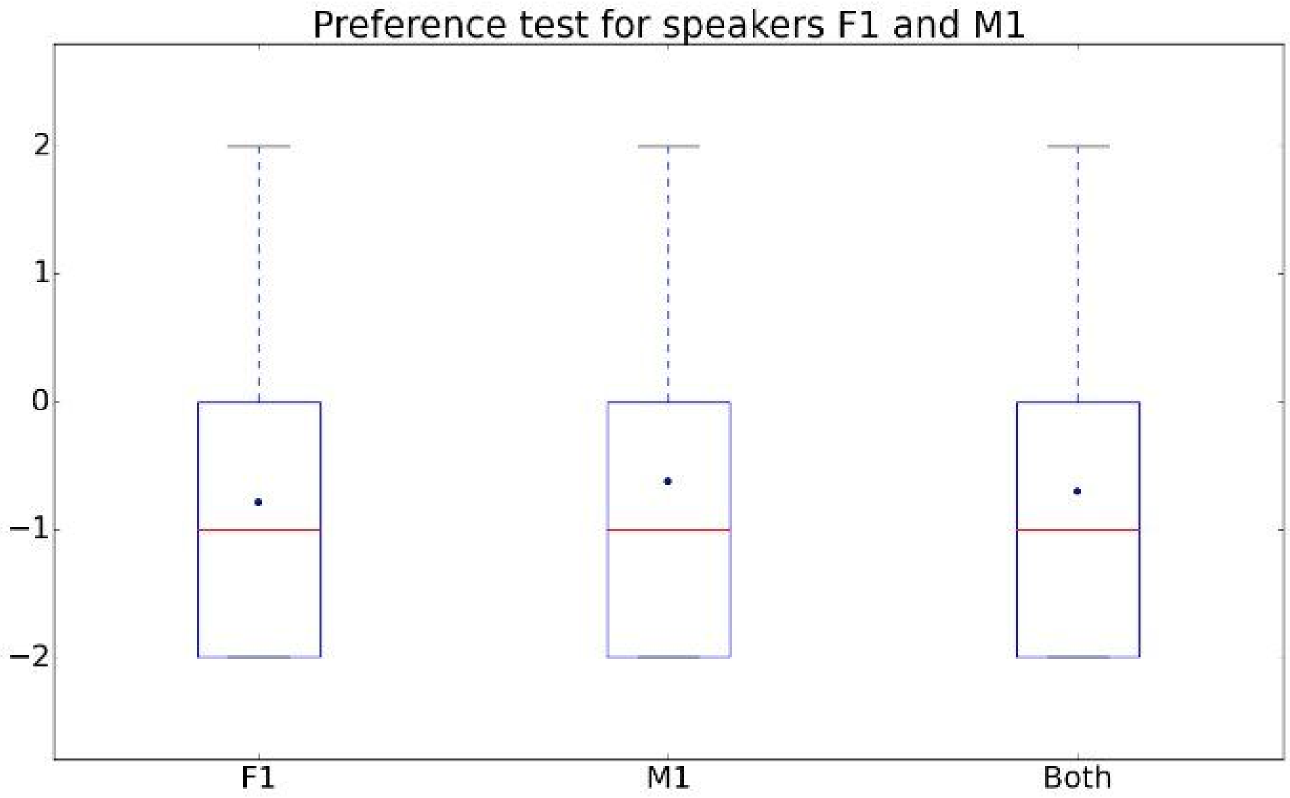
Recurrent Networks: LSTM



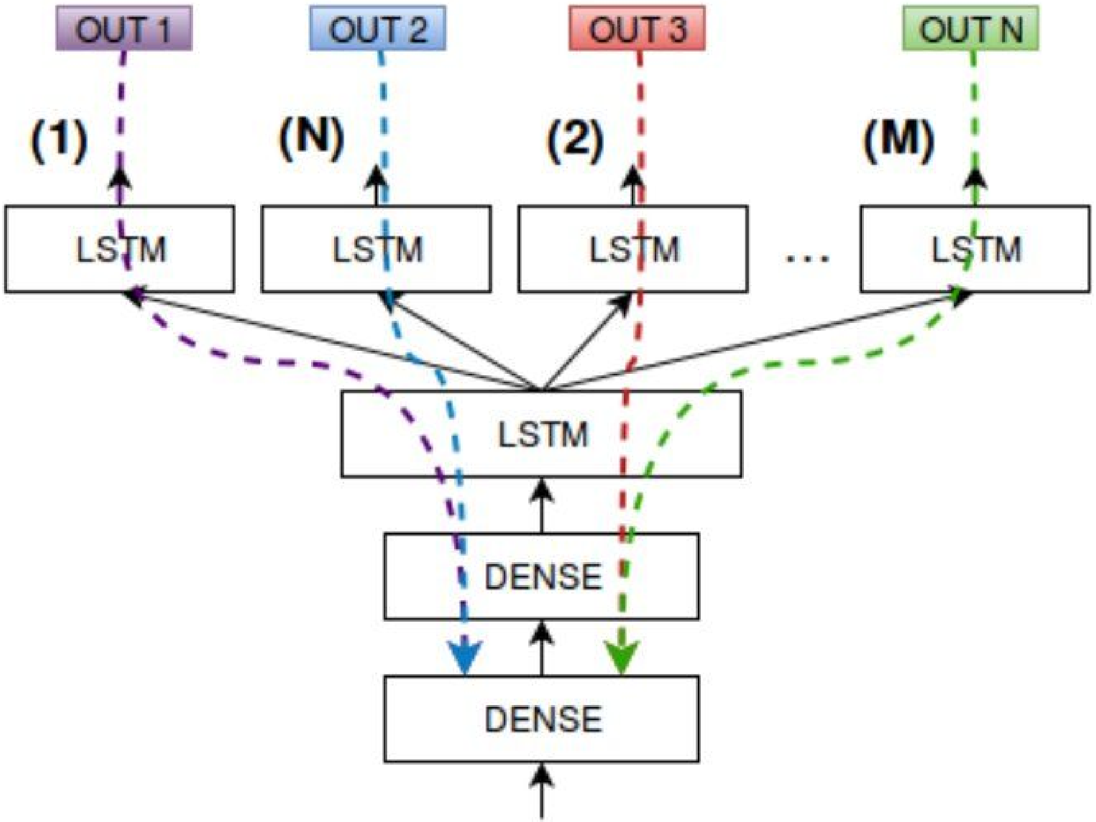
Multi-speaker



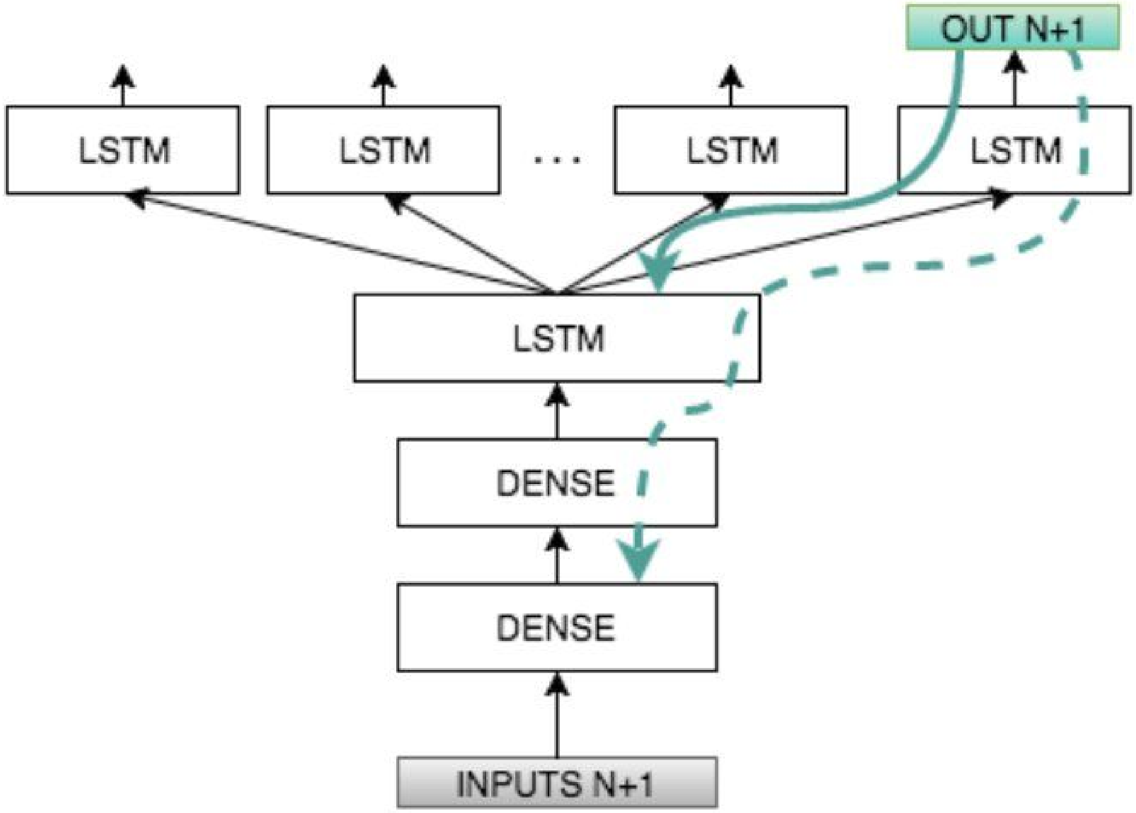
Multi-speaker



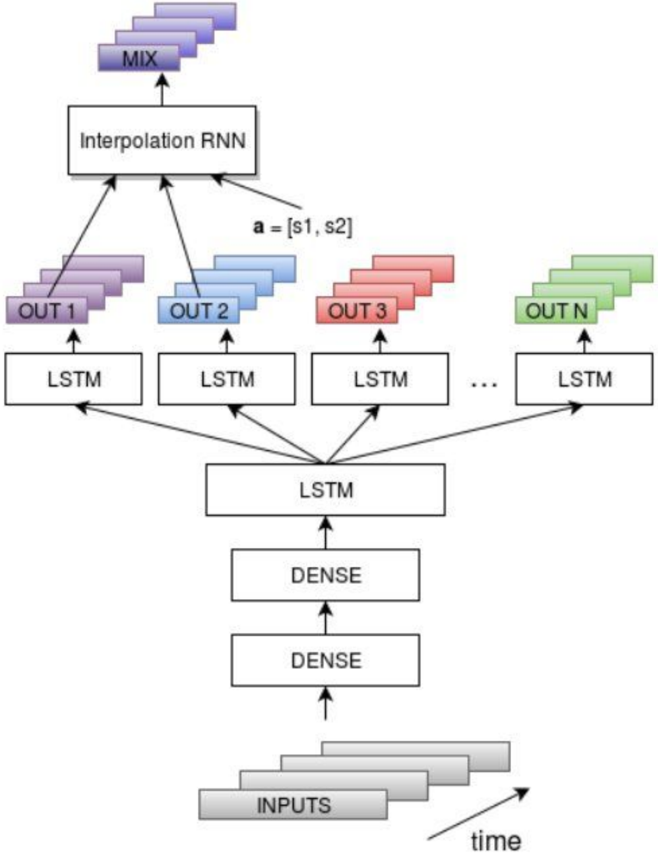
Multi-speaker



Adaptation to new speaker



Speaker interpolation



[0 1]()

[0.25 0.75]()

[0.5 0.5]()

[0.75 0.25]()

[1 0]()

References

Statistical parametric speech synthesis: from HMM to LSTM-RNN.
Heiga Zen, Google
http://rtthss2015.talp.cat/

Deep learning applied to Speech Synthesis, Msc Thesis
Santiago Pascual, UPC