

(Text to) Speech Synthesis

Yannis Stylianou

Computer Science Dept., University of Crete

Speech Synthesis – common past issues

- The spelling of words doesn't match their sound
 - Pronunciation rules + an exceptions dictionary
- Some words have multiple meanings + sounds
 - Must guess which is the correct sound
- Simplistic speech models sound mechanical
 - Can use extracts from real speech
- Speech sounds are influenced by adjacent phonemes
 - Use phoneme pairs from real speech
- Important words must be slightly louder
 - Must try to understand the text
- Voice pitch and talking speed must vary smoothly throughout a sentence
 - Must be able to change pitch and speed without affecting formant frequencies

Speech Synthesis – current state

- Is the standard Text-to-Speech synthesis problem solved?

YES!

- Which problem is still around:

Controllability

Outline

- Short overview
- Current concatenative systems – in a nutshell
- Statistical models – Regression
- Quick review of DNNs – a fast tour
- Neural TTS – sequence-to-sequence models
- Current Issues
- Applications
- Learning more ...

Outline

- Short overview
- Current concatenative systems – in a nutshell
- Statistical models – Regression
- Quick review of DNNs – a fast tour
- Neural TTS – sequence-to-sequence models
- Current Issues
- Applications
- Learning more ...

Definitions

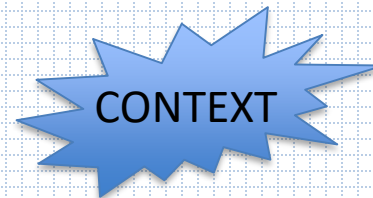
- Speech synthesis is the artificial production of human speech
(Wikipedia)
- Text-to-Speech (TTS) refers to the conversion of text to intelligible, natural and expressive speech (it has a history of over 50 years)

Text-to-Speech









- Text-to-Speech (TTS) refers to the conversion of text to intelligible, natural and expressive speech
- An ill-posed problem:
 - **Text** - a narrow band information – **to Speech** - a wide band information
- A solution: record all the words and just play them back

read it!

I have read it!



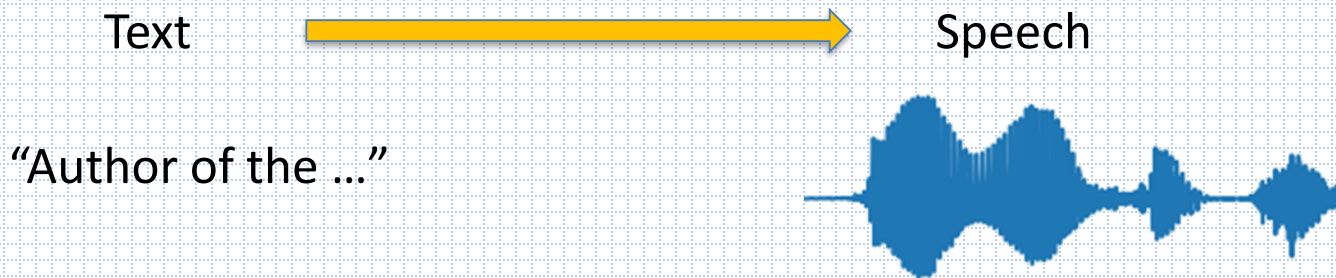
Text-to-Speech – the path so far

- Formant synthesizers 
- Diphones 
- Unit selection 
- Statistical Parametric    (Hybrid)
- Neural based  (wavenet vocoder)  (Tacotron)

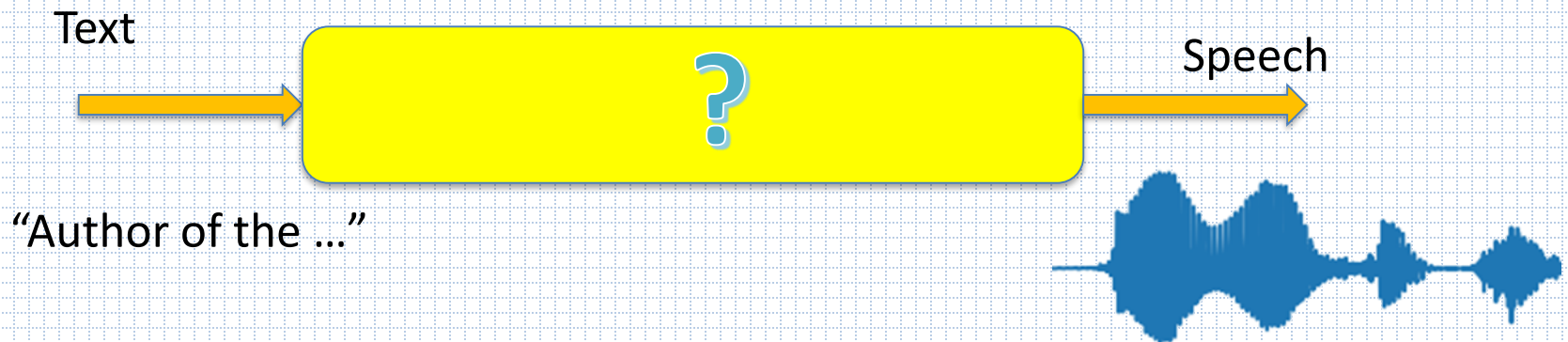
The first 3 audio files are from <https://www.ims.uni-stuttgart.de/institut/mitarbeiter/moehler/synthspeech/#english>

The last audio file (Tacotron) is from <https://google.github.io/tacotron/>

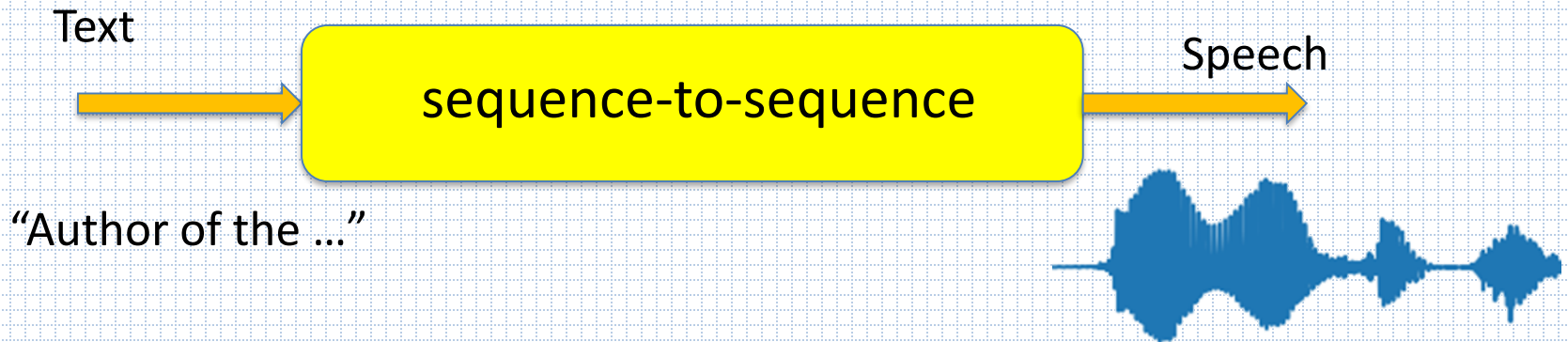
Text-to-Speech (as simple as that)



➤ End-to-end speech synthesis



Text-to-Speech ... a mapping problem

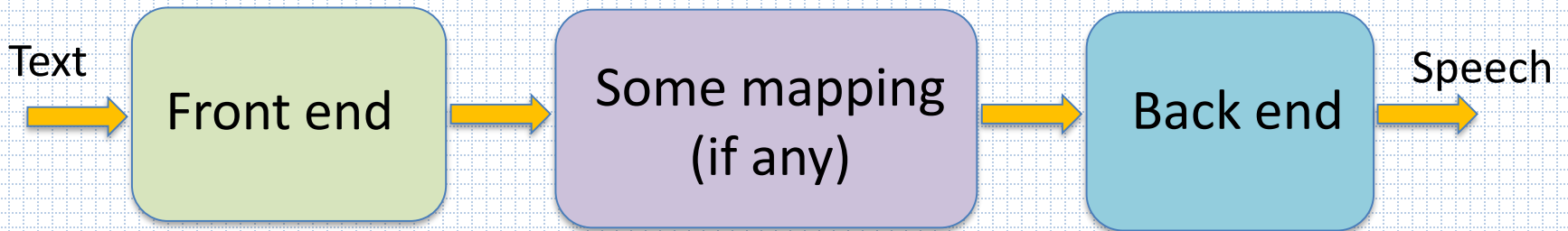
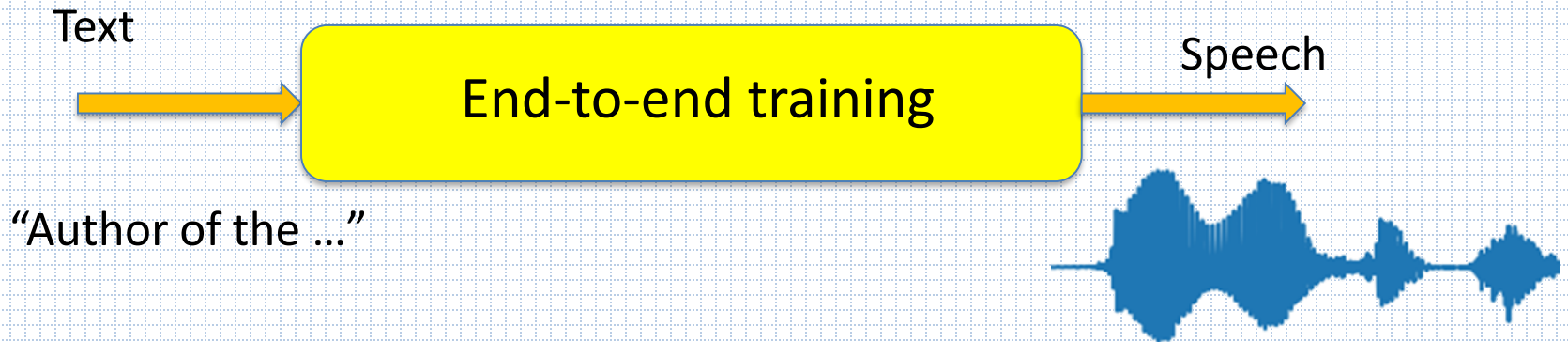


□ Options:

- Characters to samples
- Phonemes to speech features and then to samples
- Linguistic features to speech features and then to samples
- Linguistic features to samples

A sequence
to
sequence
problem

Text-to-Speech: the general framework



Outline

- Short overview
- **Current concatenative systems – in a nutshell**
- Statistical models – Regression
- Quick review of DNNs – a fast tour
- Neural TTS – sequence-to-sequence models
- Current Issues
- Applications
- Learning more ...

Features from text - linguistics

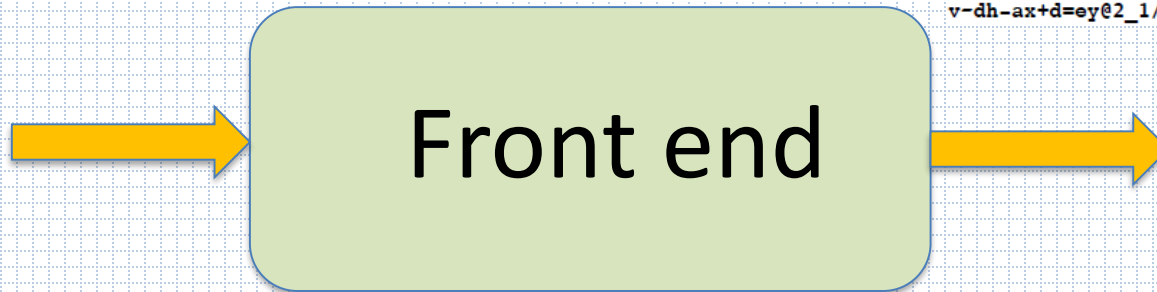
“Author of the ...”

```
sil-sil-sil+ao=th@x_x/A:0_0_0/B:x-x-x@x-x&x-x#x-x$...
sil-sil-ao+th=er@1_2/A:0_0_0/B:1-1-2@1-2&1-7#1-4$...
sil-ao-th+er=ah@2_1/A:0_0_0/B:1-1-2@1-2&1-7#1-4$...
ao-th-er+ah=v@1_1/A:1_1_2/B:0-0-1@2-1&2-6#1-4$...
th-er-ah+v=dh@1_2/A:0_0_1/B:1-0-2@1-1&3-5#1-3$...
er-ah-v+dh=ax@2_1/A:0_0_1/B:1-0-2@1-1&3-5#1-3$...
ah-v-dh+ax=d@1_2/A:1_0_2/B:0-0-2@1-1&4-4#2-3$...
v-dh-ax+d=ey@2_1/A:1_0_2/B:0-0-2@1-1&4-4#2-3$...
```

Features from text - linguistics

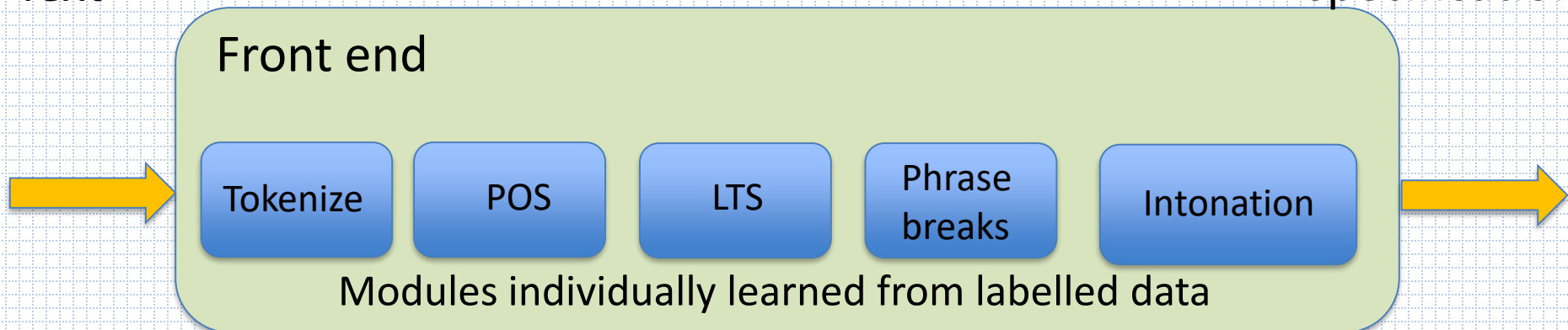
“Author of the ...”

```
sil-sil-sil+ao=thex_x/A:0_0_0/B:x-x-x@x-x&x-x#x-x$...  
sil-sil-ao+th=er@1_2/A:0_0_0/B:1-1-2@1-2&1-7#1-4$...  
sil-ao-th+er=ah@2_1/A:0_0_0/B:1-1-2@1-2&1-7#1-4$...  
ao-th-er+ah=v@1_1/A:1_1_2/B:0-0-1@2-1&2-6#1-4$...  
th-er-ah+v=dh@1_2/A:0_0_1/B:1-0-2@1-1&3-5#1-3$...  
er-ah-v+dh=ax@2_1/A:0_0_1/B:1-0-2@1-1&3-5#1-3$...  
ah-v-dh+ax=d@1_2/A:1_0_2/B:0-0-2@1-1&4-4#2-3$...  
v-dh-ax+d=ey@2_1/A:1_0_2/B:0-0-2@1-1&4-4#2-3$...
```



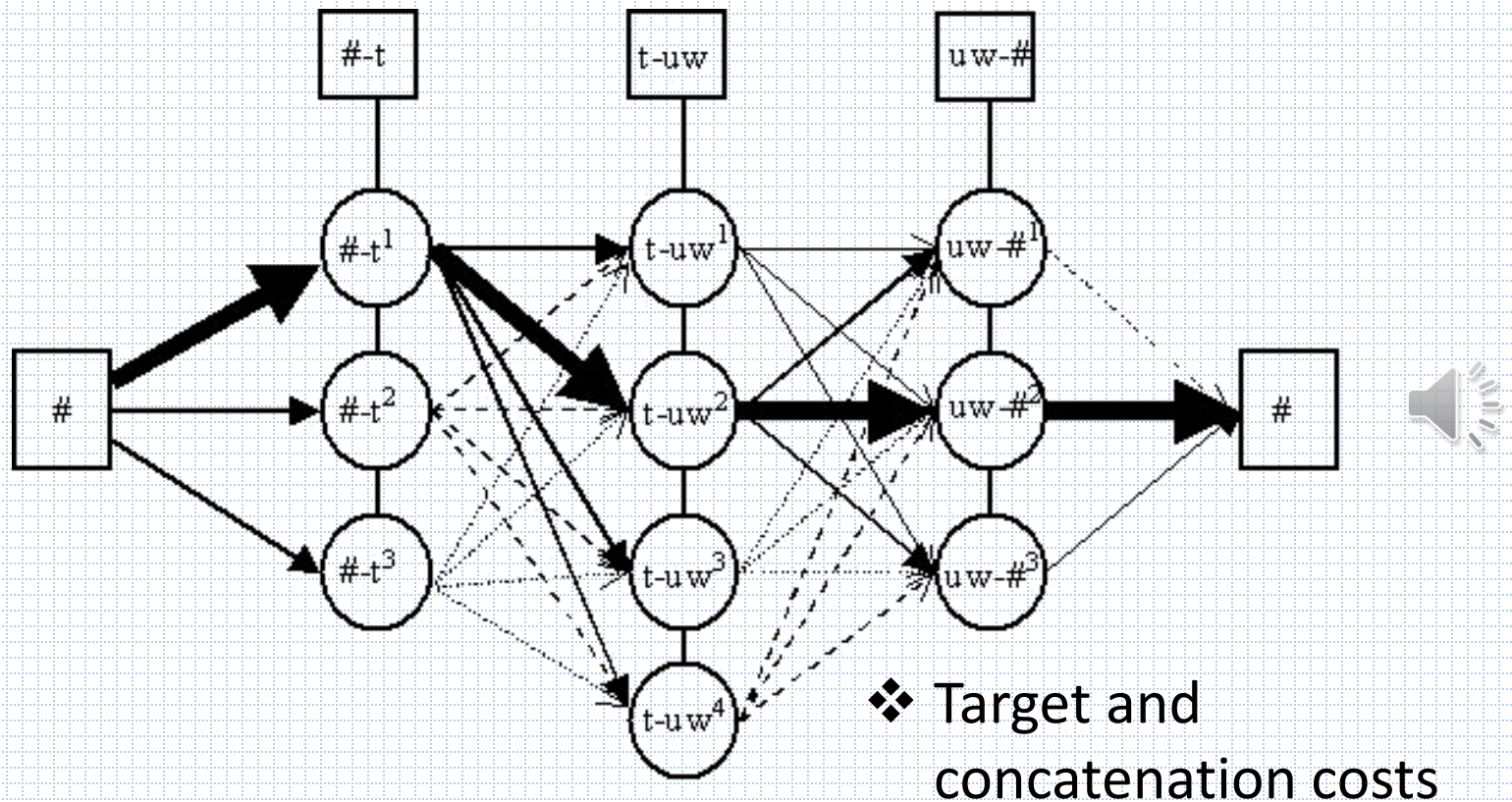
Text

Linguistic
specification



Concatenative systems (pure)

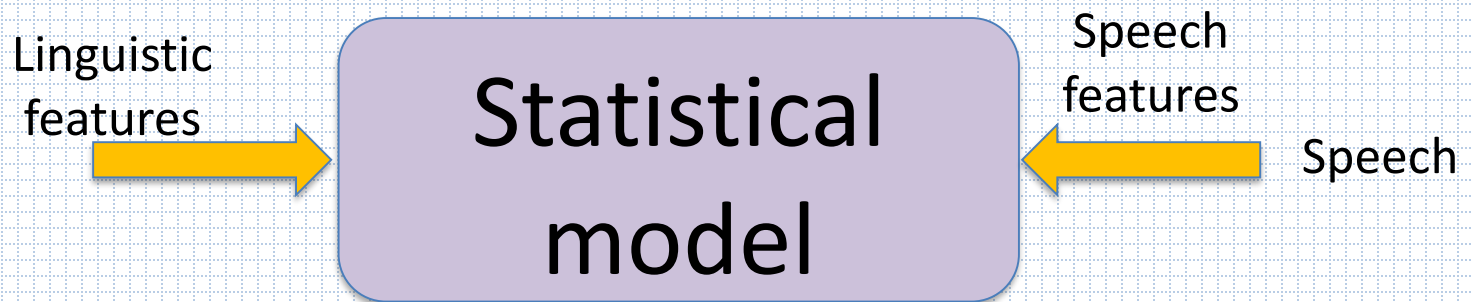
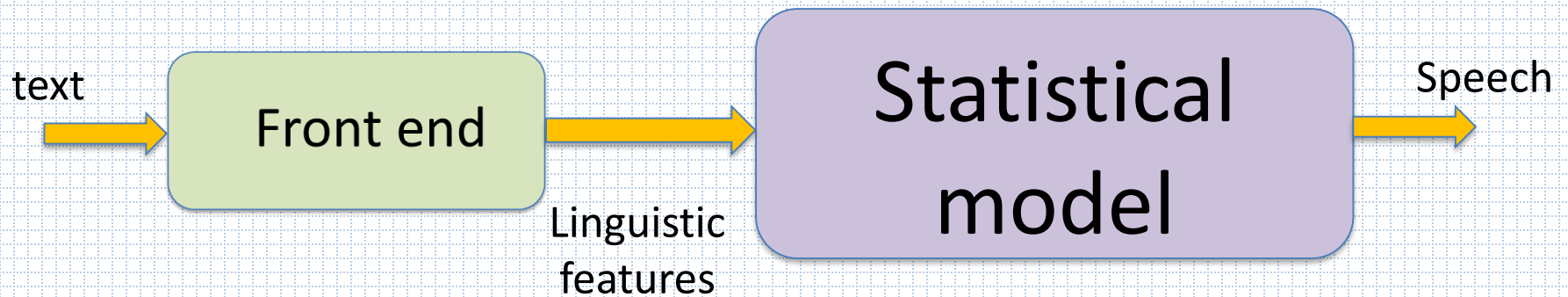
- From linguistic features to units (samples)



Outline

- Short overview
- Current concatenative systems – in a nutshell
- **Statistical models – Regression**
- Quick review of DNNs – a fast tour
- Neural TTS – sequence-to-sequence models
- Current Issues
- Applications
- Learning more ...

Start learning from data



STRAIGHT Analysis-Synthesis (H. Kawahara)

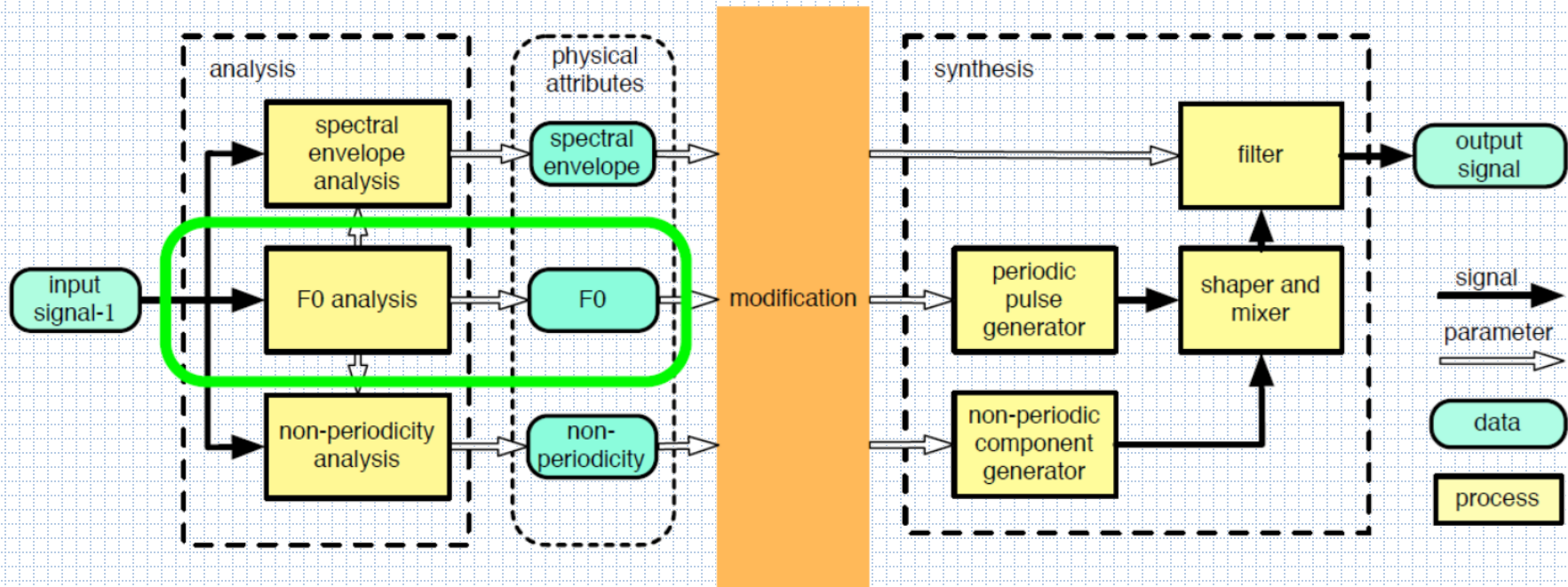
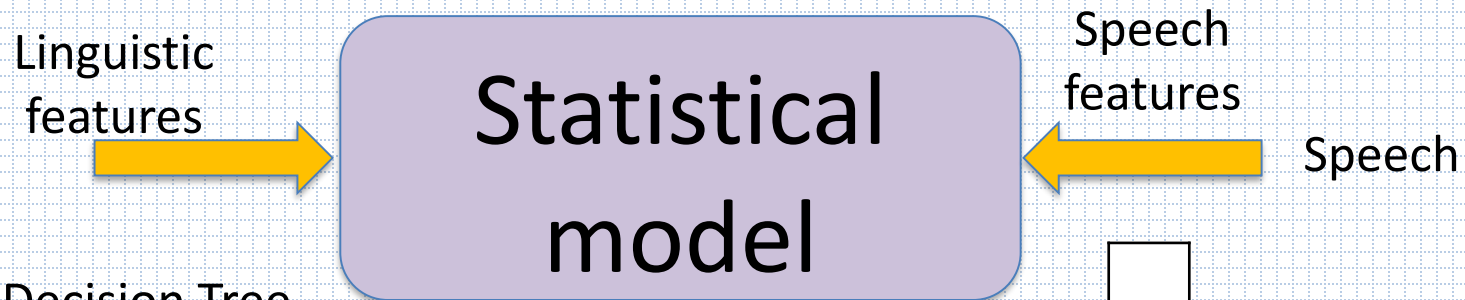
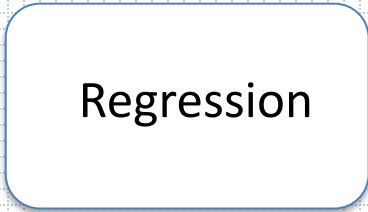
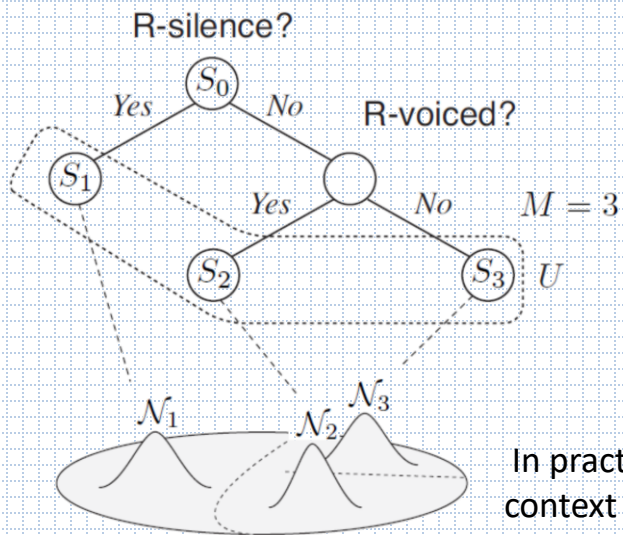


Figure from H. Kawahara

Start learning from data



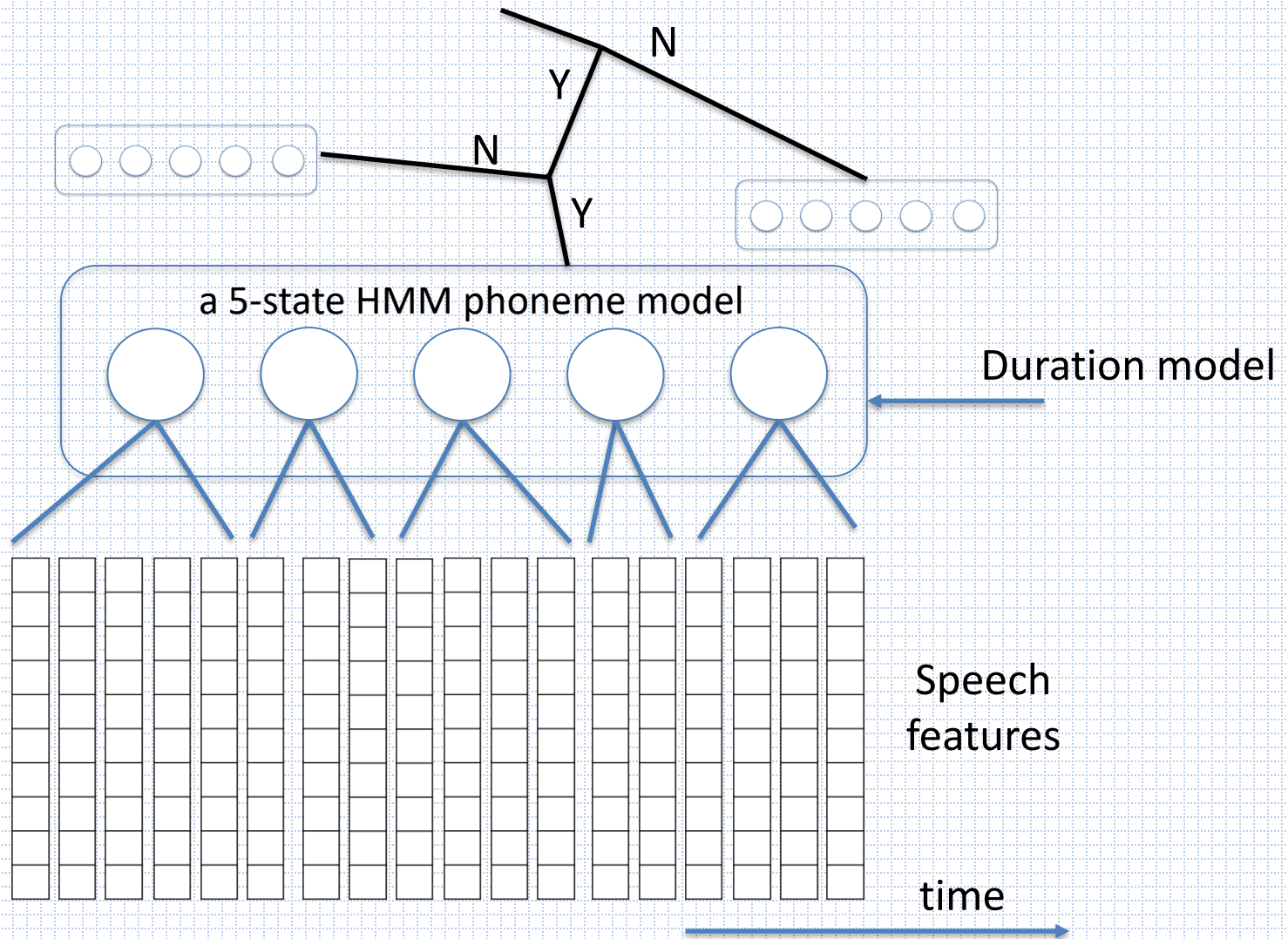
Decision Tree



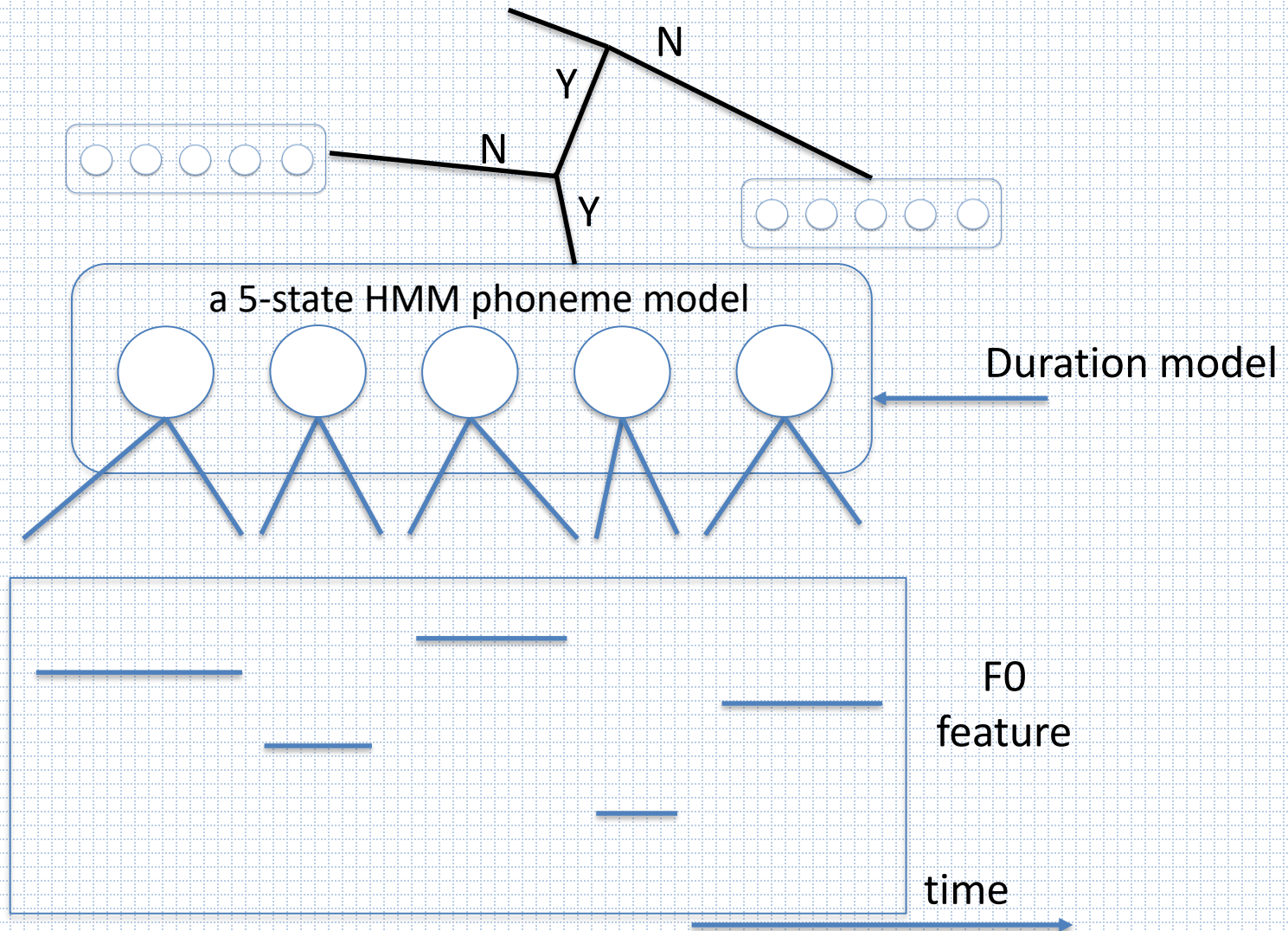
In practice we use context dependent HMMs

Figure from J. Yamagishi

Text-to-features using CART



Generating trajectories – example with F0



STRAIGHT Synthesis (H. Kawahara)

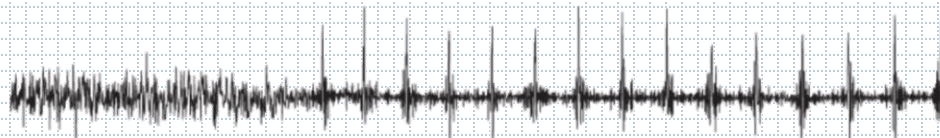
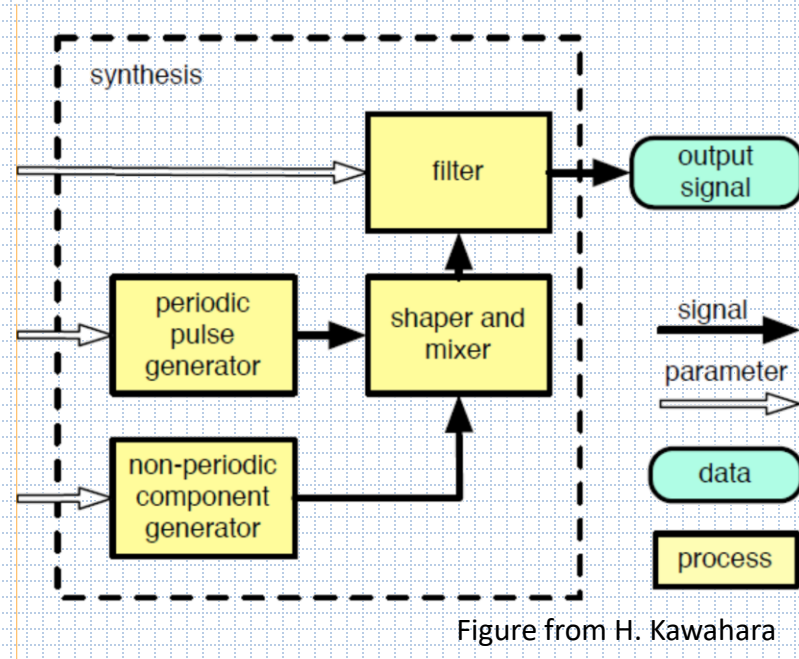
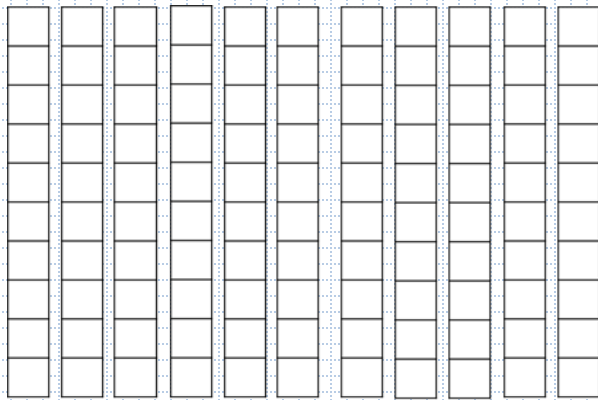
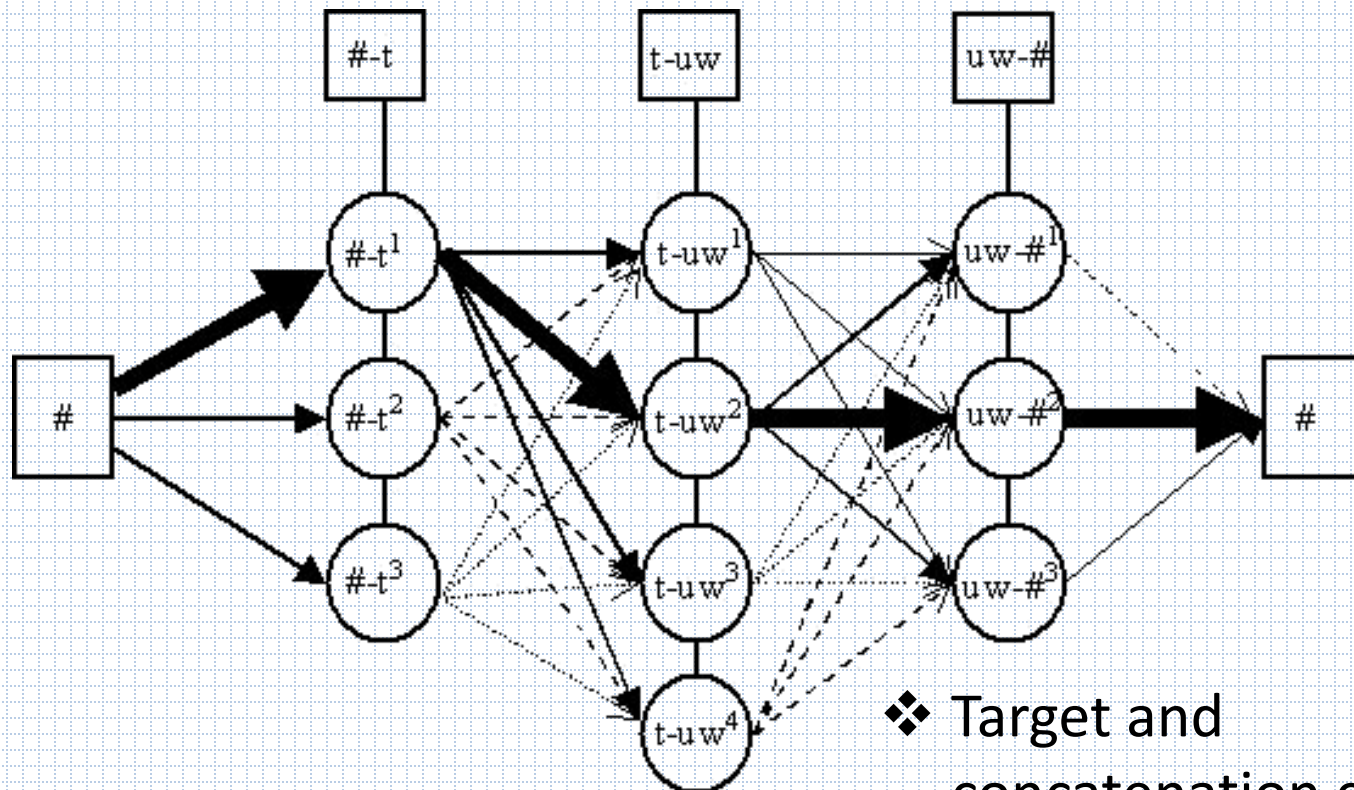


Figure from H. Zen

Concatenative systems (hybrid)

- ❑ From linguistic features to units (samples)
- ❑ Prediction of targets using HMMs

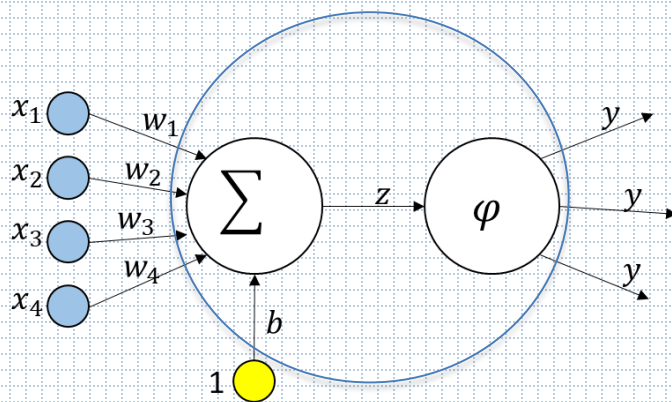
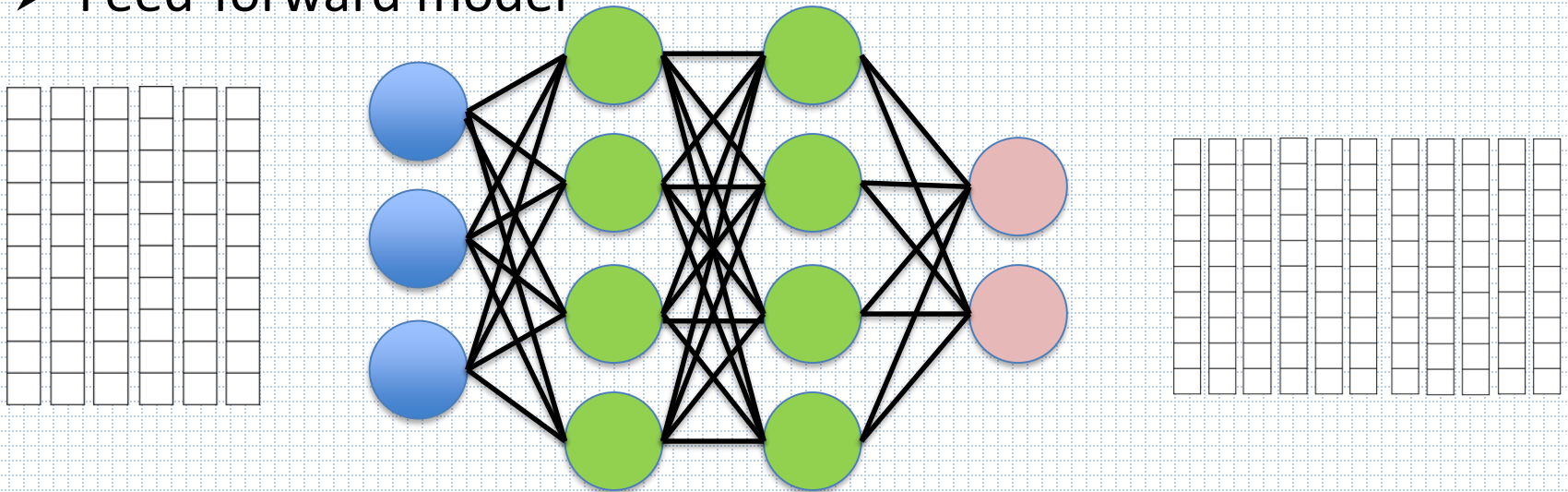


Outline

- Short overview
- Current concatenative systems – in a nutshell
- Statistical models – Regression
- **Quick review of DNNs – a fast tour**
- Neural TTS – sequence-to-sequence models
- Current Issues
- Applications
- Learning more ...

Towards neural (based) TTS - DNN

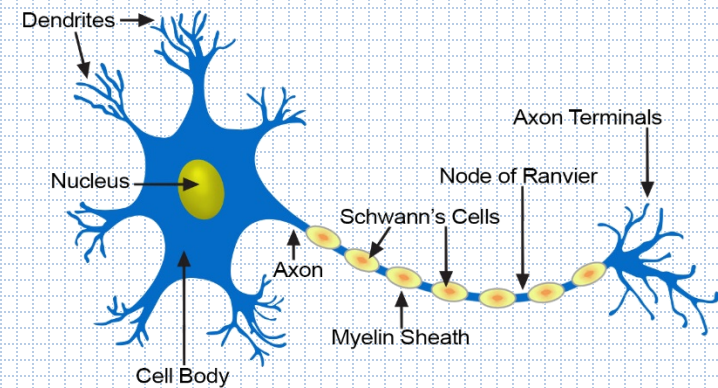
➤ Feed-forward model



$$z = \sum_i w_i x_i + b$$

$$y = \phi(z) = \begin{cases} 0, & z < 0 \\ 1, & z \geq 0 \end{cases}$$

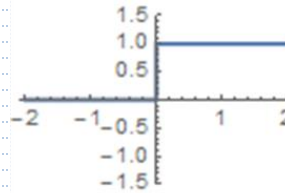
Structure of a Typical Neuron



Activation functions

Step function

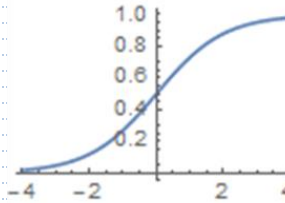
$$\varphi(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$$



$$\frac{d\varphi(x)}{dx} = \begin{cases} 0, & x < 0 \\ 0, & x > 0 \end{cases}$$

Sigmoid functions
 $\sigma(x)$

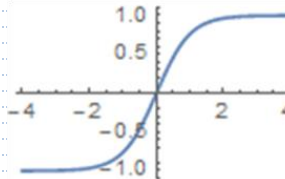
$$\varphi(x) = \frac{1}{1 + e^{-x}}$$



$$\frac{d\varphi(x)}{dx} = (1 - \varphi(x))\varphi(x)$$

Hyperbolic tangent
 $\tanh(x)$

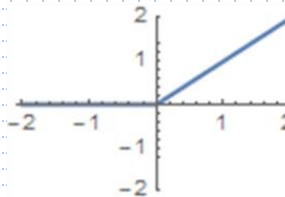
$$\varphi(x) = \frac{2}{1 + e^{-2x}} - 1$$



$$\frac{d\varphi(x)}{dx} = 1 - \varphi(x)^2$$

Rectified linear unit
ReLU

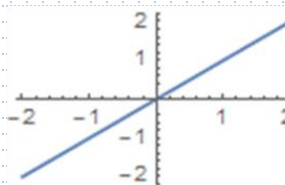
$$\varphi(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$



$$\frac{d\varphi(x)}{dx} = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

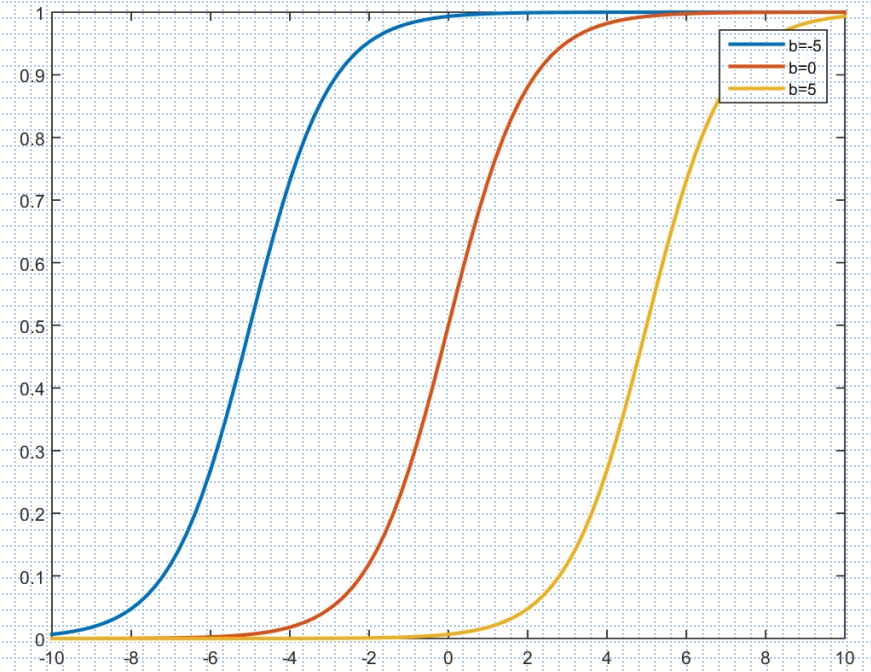
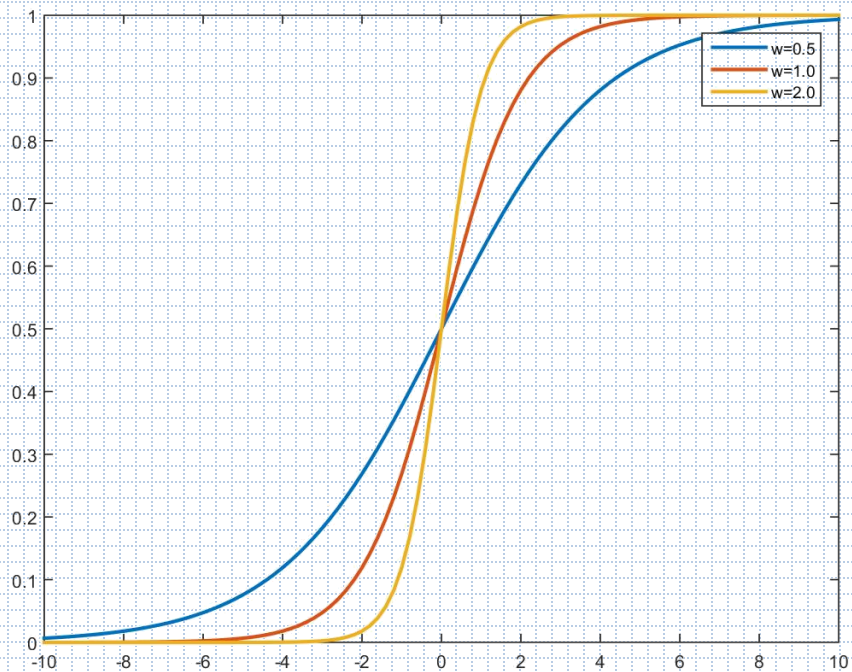
Linear function

$$\varphi(x) = x$$

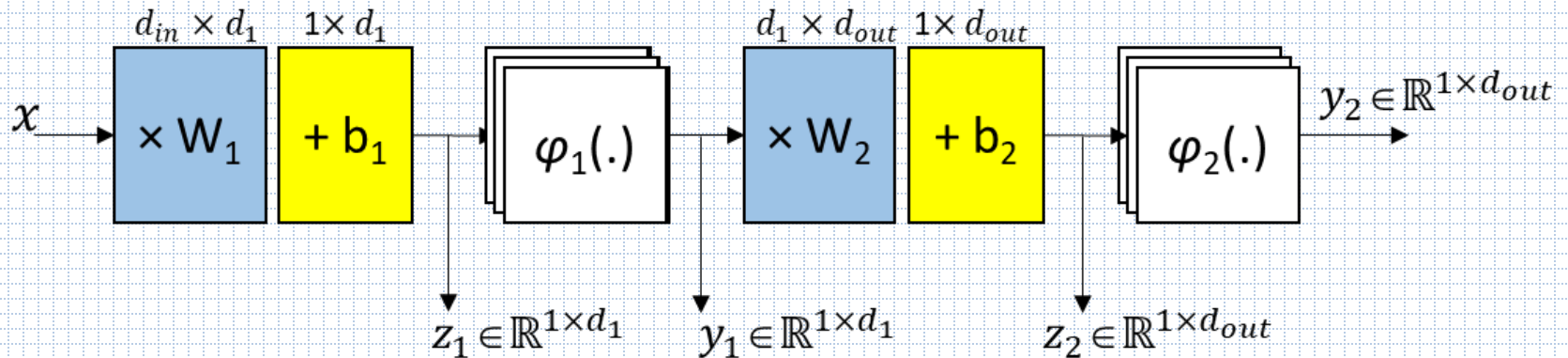


$$\frac{d\varphi(x)}{dx} = 1$$

Sigmoids – weights and bias



Usual representation of DNNs



$$z_1 = xW_1 + b_1$$

$$y_1 = \varphi_1(z_1)$$

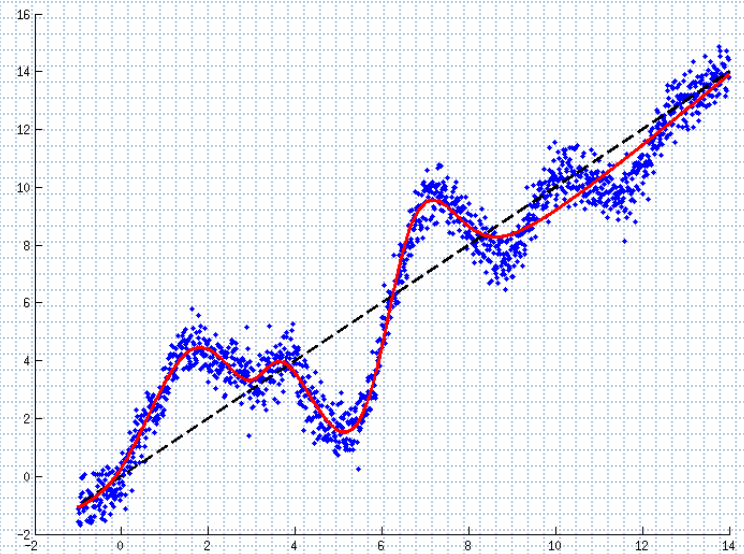
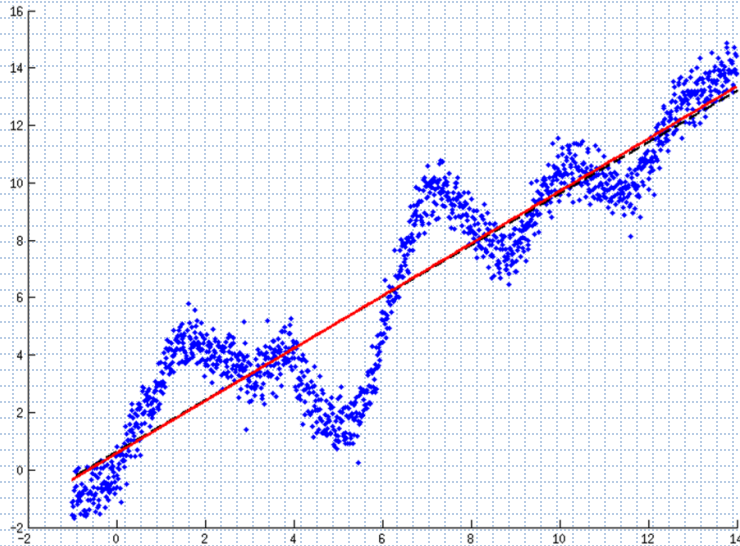
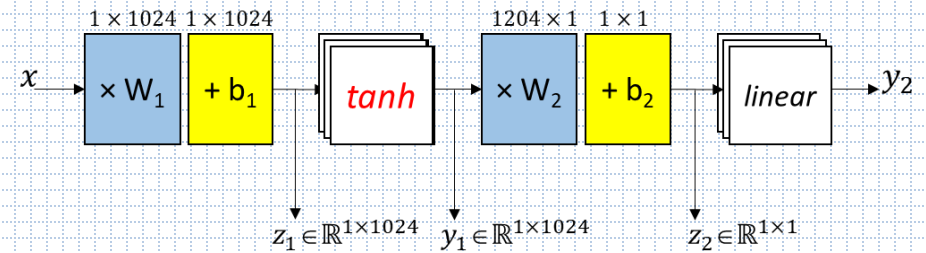
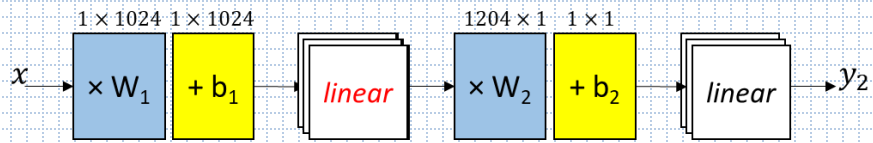
$$z_2 = y_1W_2 + b_2$$

$$y_2 = \varphi_2(z_2)$$

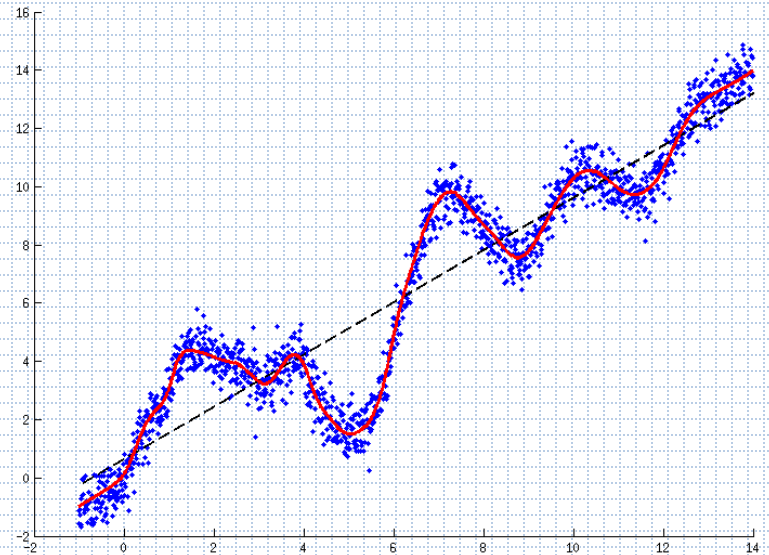
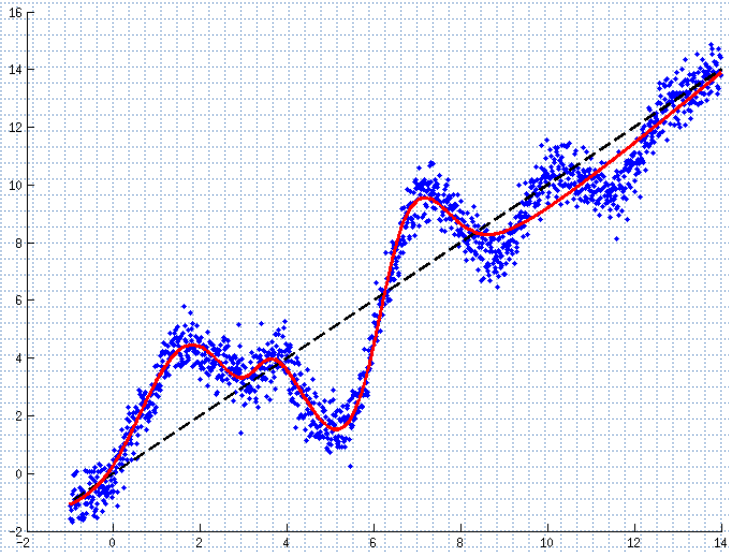
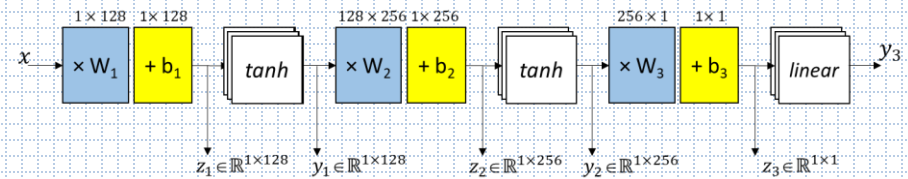
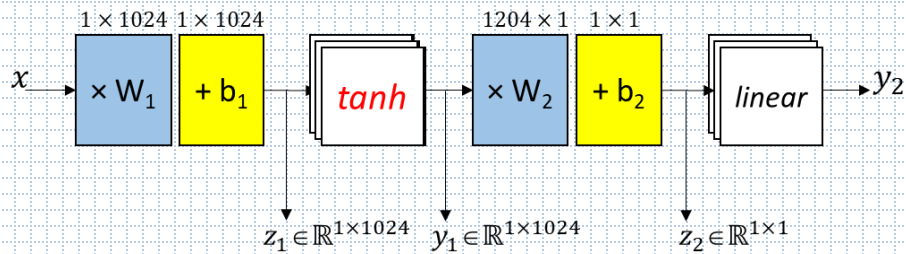
○ Softmax activation:

$$y_l(k) = \frac{e^{z_l(k)}}{\sum_{i=1}^{d_l} e^{z_l(i)}}$$

Linear/Non-Linear Neural Networks



Deep/Non-Deep Neural Networks

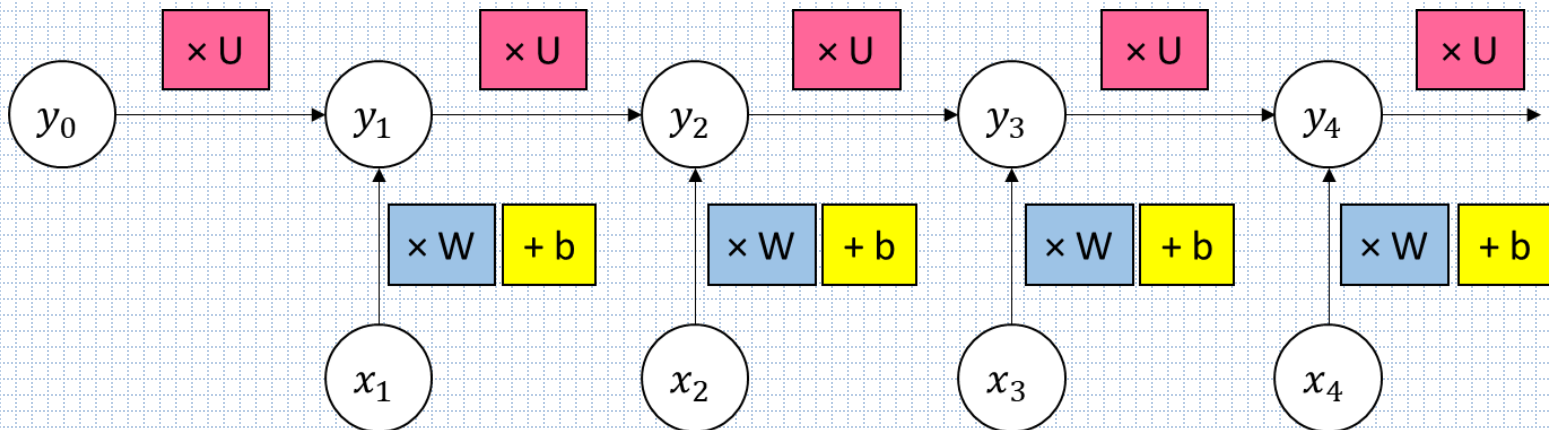
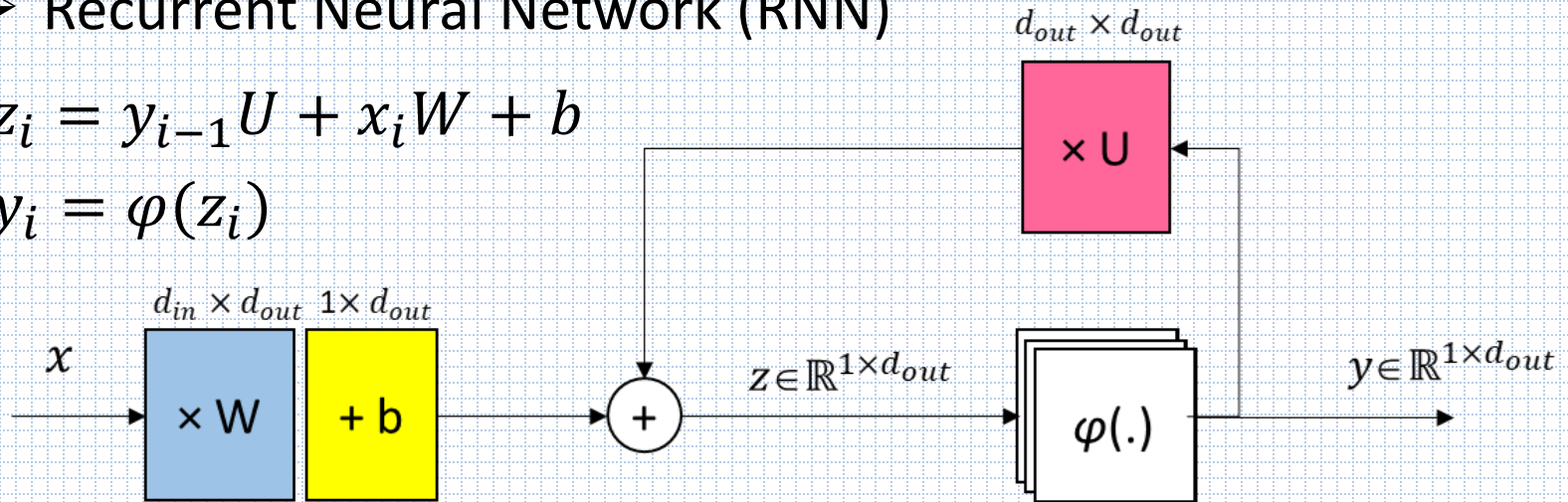


Other types of Deep Neural Networks

➤ Recurrent Neural Network (RNN)

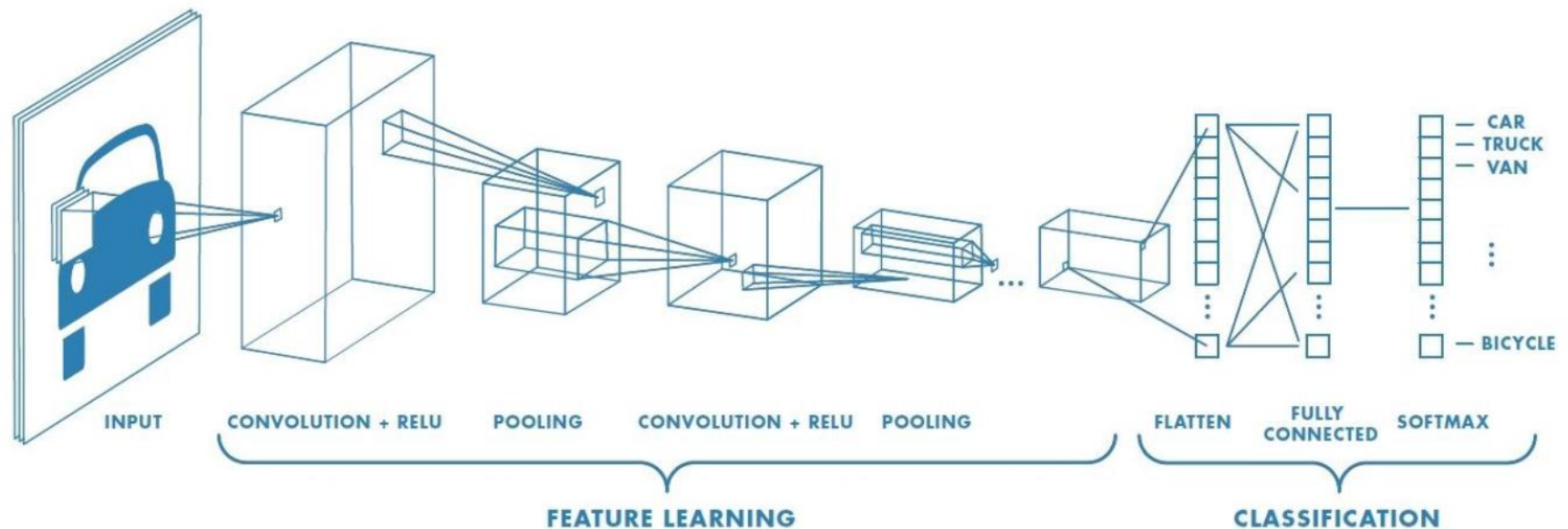
$$z_i = y_{i-1}U + x_iW + b$$

$$y_i = \varphi(z_i)$$



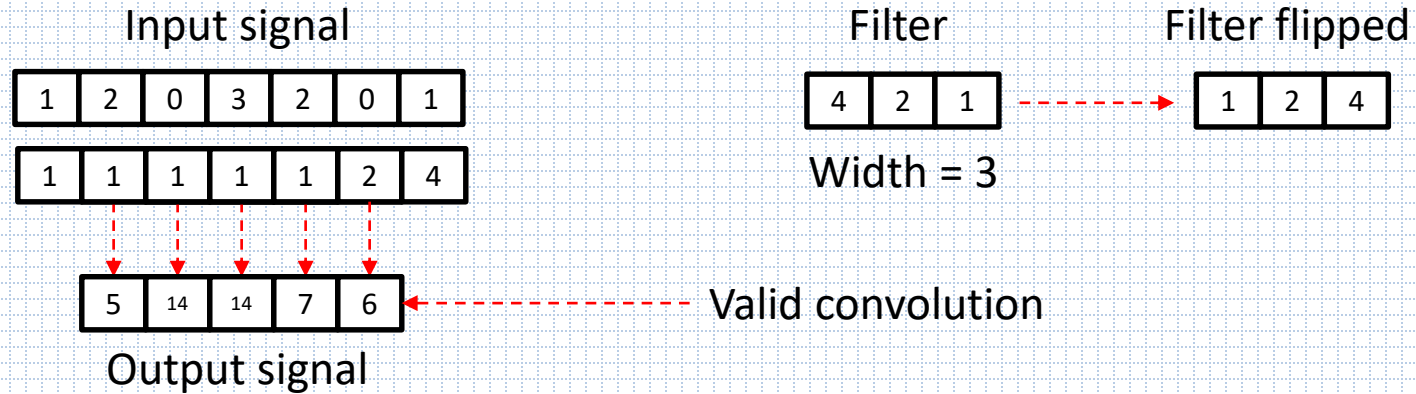
Other types of Deep Neural Networks

➤ Convolutional Neural Network (CNN)

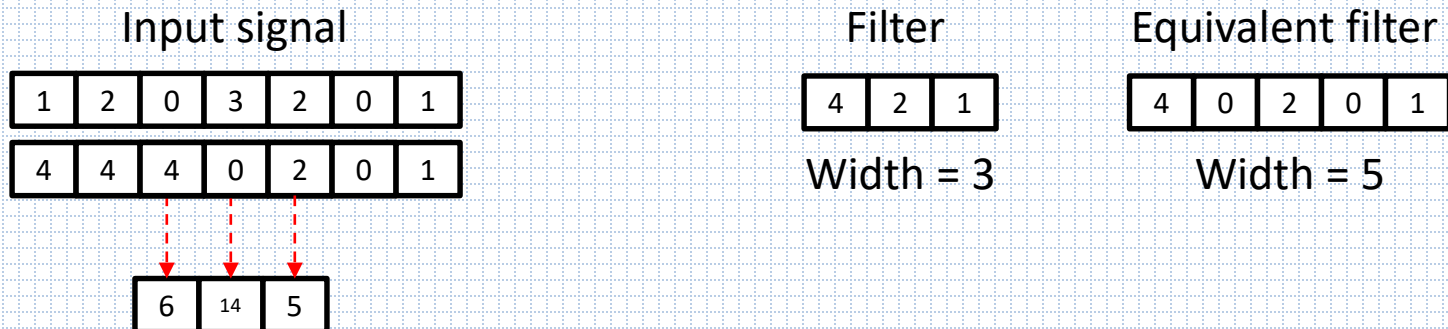


1-D convolutions

- Standard convolution



- Dilated convolution (dilation=2)



Outline

- Short overview
- Current concatenative systems – in a nutshell
- Statistical models – Regression
- Quick review of DNNs – a fast tour
- **Neural TTS – sequence-to-sequence models**
- Current Issues
- Applications
- Learning more ...

Going back to our problem: TTS (with DNNs)

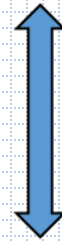
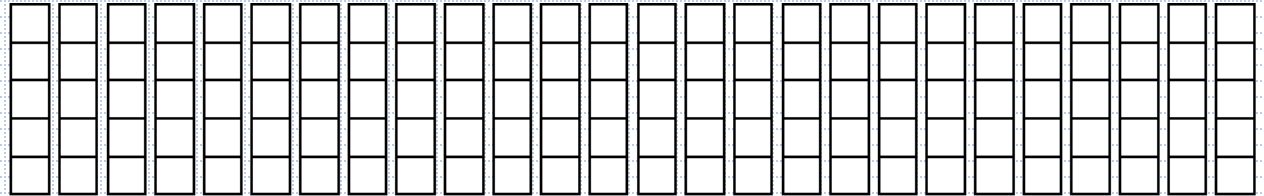
- Features encoded: context-dependent phone to a vector of binary features

```
sil-sil-sil+ao=th@x_x/A:0_0_0/B:x-x-x@x-x&x-x#x-x$. . .
sil-sil-ao+th=er@1_2/A:0_0_0/B:1-1-2@1-2&1-7#1-4$. . .
sil-ao-th+er=ah@2_1/A:0_0_0/B:1-1-2@1-2&1-7#1-4$. . .
ao-th-er+ah=v@1_1/A:1_1_2/B:0-0-1@2-1&2-6#1-4$. . .
th-er-ah+v=dh@1_2/A:0_0_1/B:1-0-2@1-1&3-5#1-3$. . .
er-ah-v+dh=ax@2_1/A:0_0_1/B:1-0-2@1-1&3-5#1-3$. . .
ah-v-dh+ax=d@1_2/A:1_0_2/B:0-0-2@1-1&4-4#2-3$. . .
v-dh-ax+d=ey@2_1/A:1_0_2/B:0-0-2@1-1&4-4#2-3$. . .
```



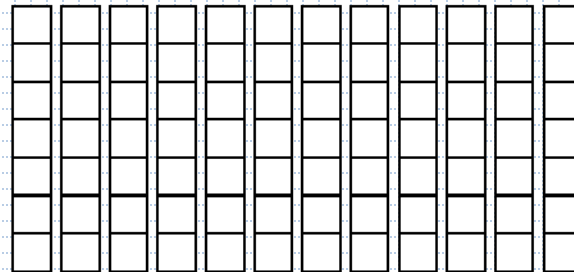
Neural TTS = a sequence-to-sequence regression

Output sequence:
speech features



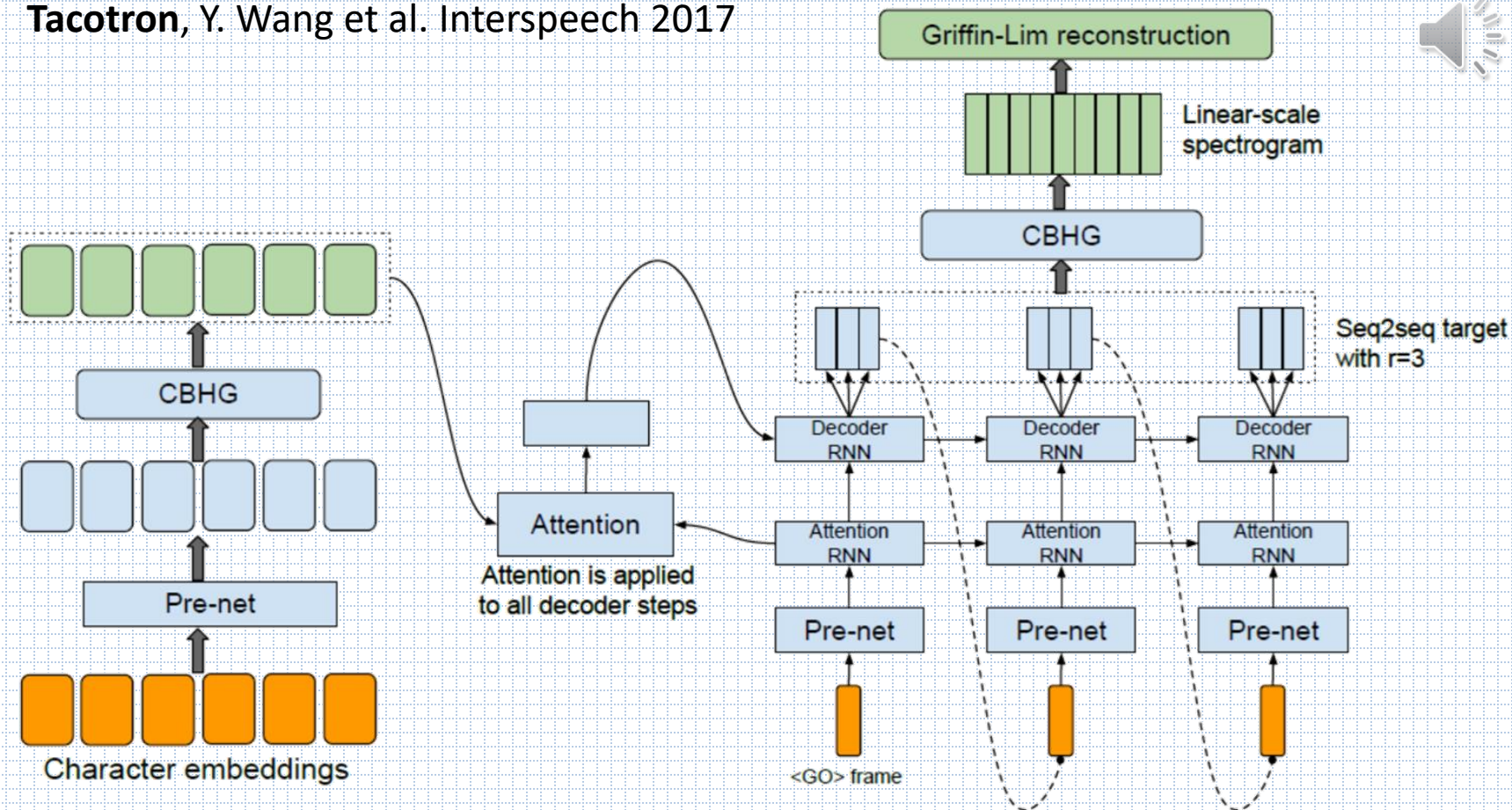
Different lengths, because of
different clock rates

Input sequence:
linguistic specification



Tacotron: a multiple sequence-to-sequence model

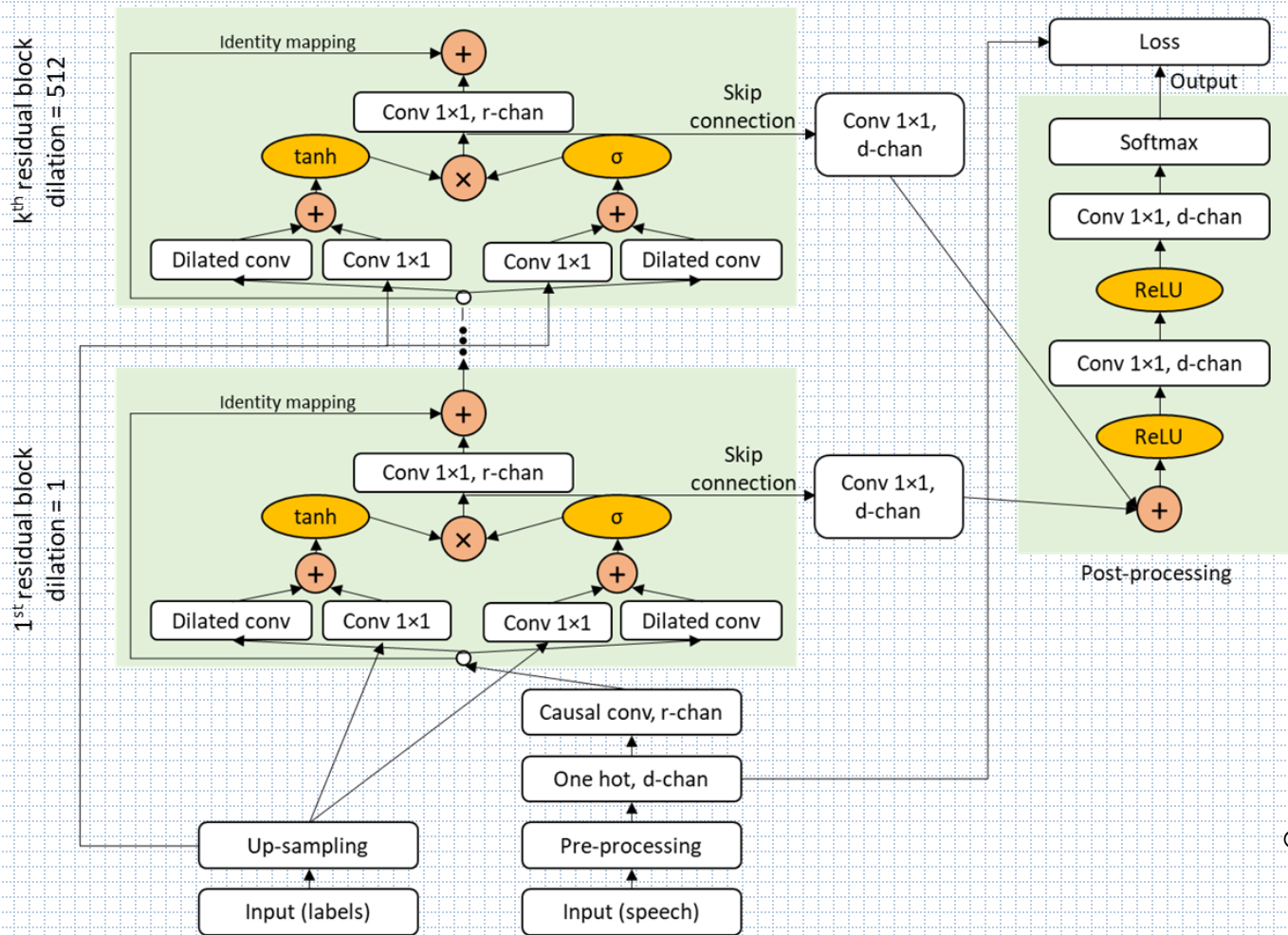
Tacotron, Y. Wang et al. Interspeech 2017



CHBG: Convolution bank – highway network – bidirectional Gated Recurrent Unit (GRU)

Wavenet

$$P(x_n | x_{n-1}, x_{n-2}, \dots, x_{n-r}, h_n)$$



Sound examples (16 kHz) [test data]:

- ❖ with natural prosody:
 - Google (40 hours)



- UoC (5 hours)



- ❖ with synthetic prosody (HMM):

- UoC (5 hours)



Outline

- Short overview
- Current concatenative systems – in a nutshell
- Statistical models – Regression
- Quick review of DNNs – a fast tour
- Neural TTS – sequence-to-sequence models
- **Current Issues**
- Applications
- Learning more ...

Speech Synthesis – current issues

➤ Robustness & running cost

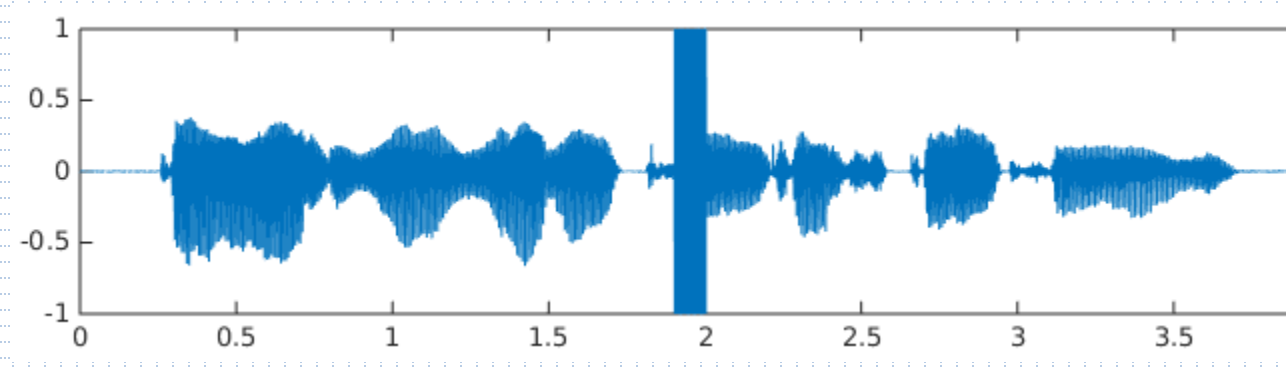
- Robust & fast front-end and back-end (Parallel Wavenet, WaveRNN, ...)
- Robust to recordings quality and quantity
- Robust training

➤ Context awareness

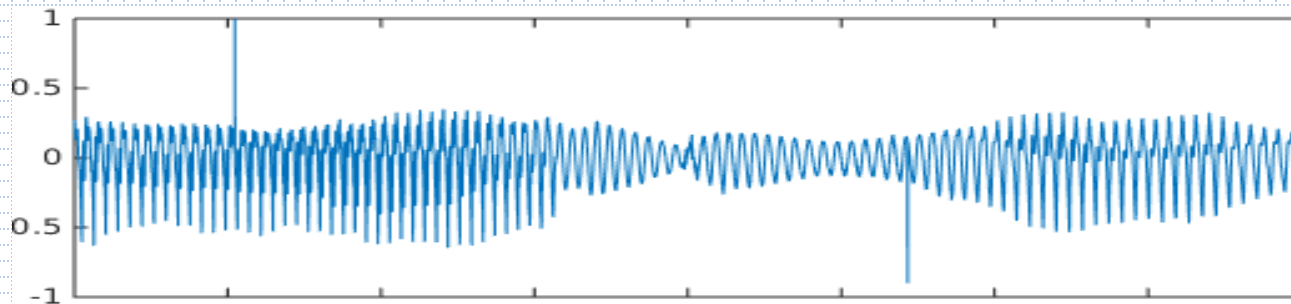
- Adaptation to user acts in dialogue (conversational TTS, style token)
- Adaptation to the listening conditions (intelligibility)

Current issues – robust back-end

- Type I artifacts, when we predict parameters of a distribution



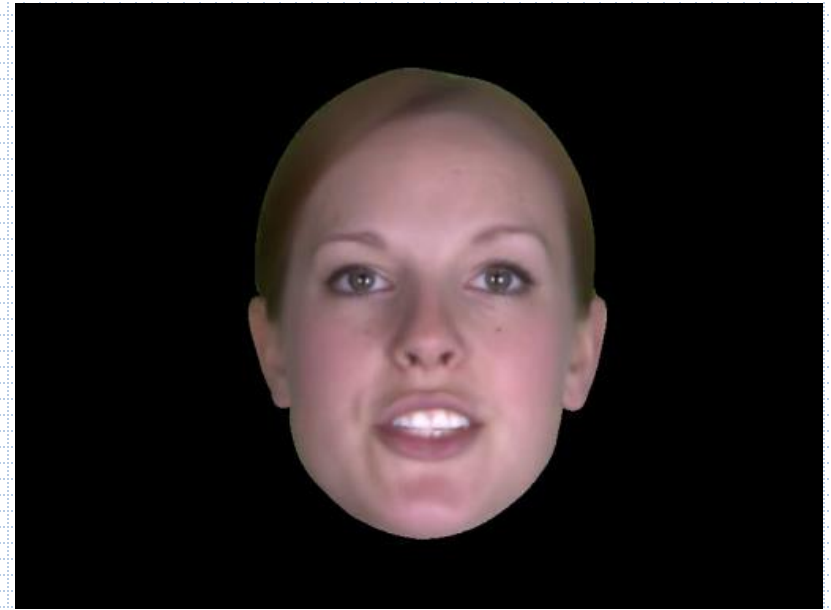
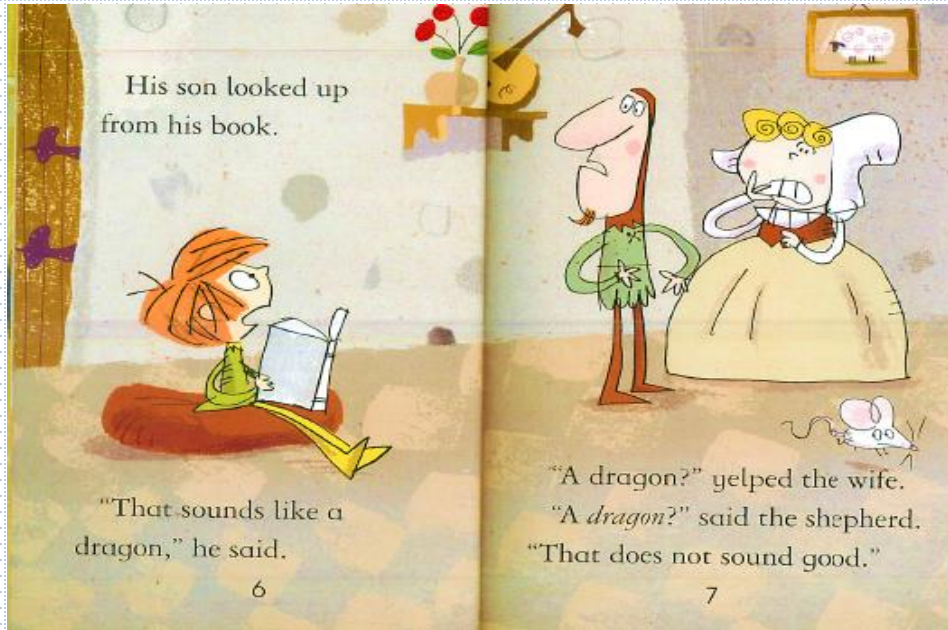
- Type II artifacts, when we produce categorical (softmax) distribution



Outline

- Short overview
- Current concatenative systems – in a nutshell
- Statistical models – Regression
- Quick review of DNNs – a fast tour
- Neural TTS – sequence-to-sequence models
- Current Issues
- **Applications**
- Learning more ...

The usual (suspect of) application



Xpressive Talk™ Toshiba Corp.



The real application: Conversational TTS



Toshiba: Statistical Dialogue System

Outline

- Short overview
- Current concatenative systems – in a nutshell
- Statistical models – Regression
- Quick review of DNNs – a fast tour
- Neural TTS – sequence-to-sequence models
- Current Issues
- Applications
- **Learning more ...**

Inclusive Neural Speech Synthesis - iNSS

Speech Processing Courses in Crete (at ICASSP22)

May 22-27, 2022

Singapore

<http://spcc.csd.uoc.gr/>

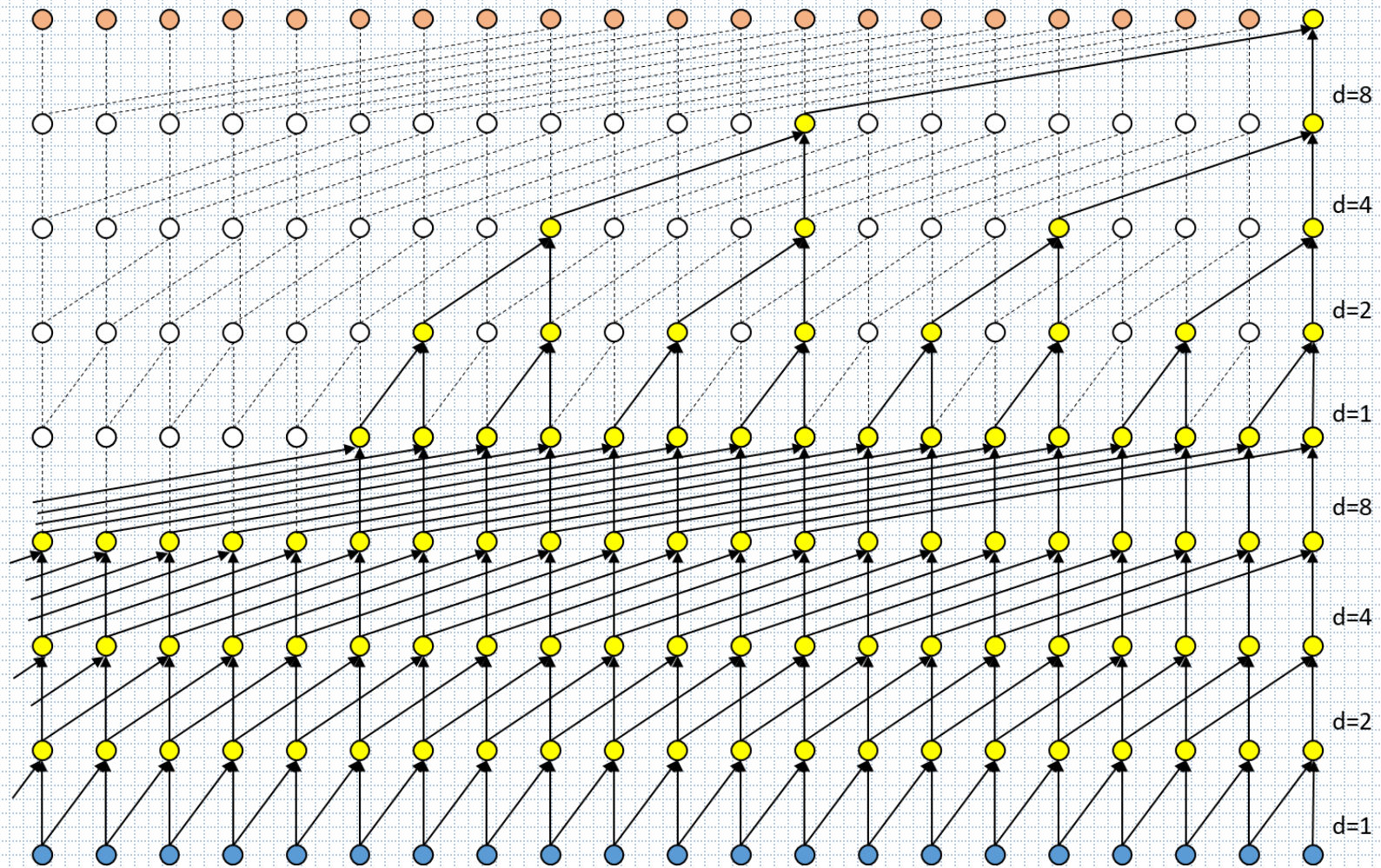
➤ Learn (with theory and hands on) about:

- ✓ Neural Front-end, Alistair Conkie, Apple USA
- ✓ Neural Acoustic Modelling (Vassilis Tsiaras, UoC, Greece)
- ✓ Neural vocoders (Junichi Yamagishi and Xin Wang, NII Japan)
- ✓ Fast speech (Malcolm Slaney, Google USA)
- ✓ Voice Banking (Youtian Chen, DeepMind, UK/China)
- ✓ Intelligibility (Yannis Stylianou, Petko Petkov and Shifas PV, Apple UK)

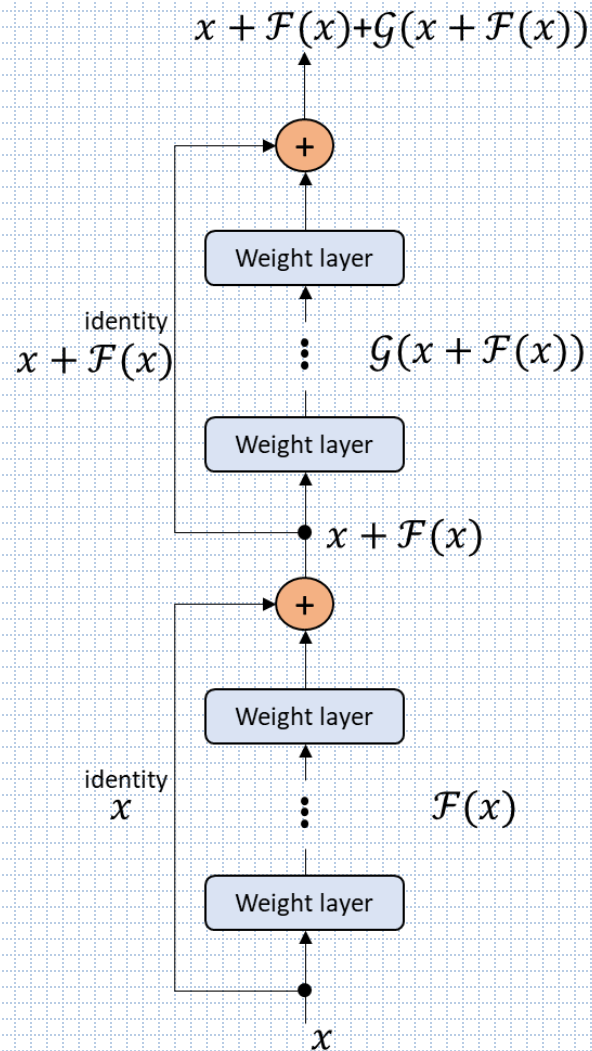
Further reading

- ✓ ISCA Synthesis Interest Group: https://synsig.org/index.php/Main_Page (look there for the SPCC videos)
- ✓ CSTR, Simon Kings' page on speech synthesis: <http://www.speech.zone/courses/speech-synthesis/>
- ✓ CMU Festvox: <http://festvox.org/>
- ✓ Google: <https://google.github.io/tacotron/>

On Wavenet – connections



On Wavenet – Residual connections



On Wavenet – gating

