Speech Signal Processing Lab

- a short tour

https://www.csd.uoc.gr/~sspl/index.html

Yannis Stylianou,

Prof. of Speech Processing,
University of Crete

Speech Signal Processing Lab

- a short tour

[*https://www.csd.uoc.gr/~sspl/index.html*](https://www.csd.uoc.gr/~sspl/index.html)

Topics:

1. General overview, about us
2. Introduction to Speech Technology
3. Text to Speech Synthesis

Speech Signal Processing Lab

- a short tour

https://www.csd.uoc.gr/~sspl/index.html

Topics:

1. General overview, about us
2. Introduction to Speech Technology
3. Text to Speech Synthesis

BioSketch

Yannis Stylianou is Professor of Speech Processing at University of Crete, in Greece and Research Manager at Apple, Cambridge UK.

From 1996 until 2001 he was with AT&T Labs Research (Murray Hill and Florham Park, NJ, USA) and until 2002 he was with Bell-Labs Lucent Technologies, in Murray Hill, NJ, USA. He is with University of Crete since 2002.

From 2013 until 2018 (July) he was Group Leader of the Speech Technology Group at Toshiba Cambridge Research Lab in Cambridge UK. He joined Apple in Aug 2018. He holds MSc and PhD from ENST-Paris on Signal Processing and he has studied Electrical Engineering at NTUA Athens Greece (1991).

He is an IEEE Fellow and an ISCA Fellow.

Speech Processing Lab

- Key people

George Kafentzis
Signal Processing

Yannis Pantazis
Signal Processing

Vassilis Tsiaras
Machine Learning

Speech Signal Processing Lab

- Key people

Ph.Ds:

1. Yannis Agiomyrgiannakis, Google UK, Altered LTD London
2. Yannis Pantazis, FORTH
3. Andre Holzapfel, Assistant Professor KTH Sweden
4. Maria Koutsogiannaki, BCBL, Spain
5. Maria Markaki, FORTH
6. George Kafentzis, UoC/CSD
7. Muhammed Shifas PV (on going)
8. Dipjyoti Paul (on going)
9. Rafael Tsirbas (on going)
10. Irene Sissamaki (to start soon)

Speech Signal Processing Lab

- Summary of topics

- ✓ Speech Processing
- ✓ Audio Processing: Music, Marine mammals
- ✓ Biomedical Signal Processing:
  - ✓ Voice function assessment
  - ✓ Phonocardiography

Speech Signal Processing Lab

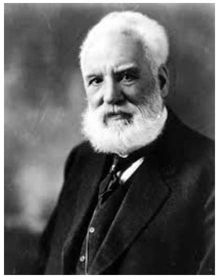
- a short tour

https://www.csd.uoc.gr/~sspl/index.html

Topics:

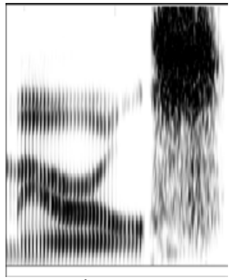
1. General overview, about us
2. Introduction to Speech Technology
3. Text to Speech Synthesis

Speech has a central position in human communication



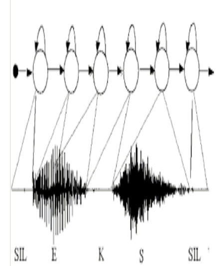
Bell (1876) discovery of telephone

Rayleigh (1900) theory of sound



Speech spectrogram (1946)

Shannon (1948) speech & language transmission



Markov chain (Baum, 1960)

Békésy (1961) frequency coding

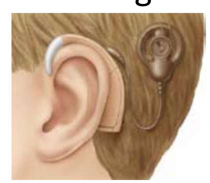
Itakura (1970) Autoregressive modelling

Turing (1950) thinking machine

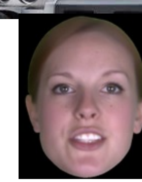
Understanding speech production and acoustics led to …

✓ Improved communication

✓ Enhanced hearing



✓ Advanced speech technologies



Google talk

Siri

Speech Image Knowledge RECAIUS™

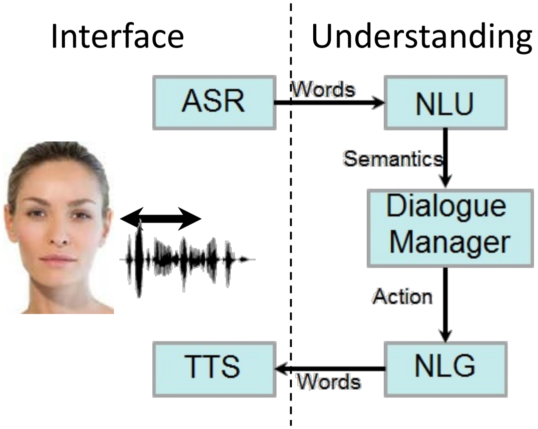
- ➢ Text-to-Speech Synthesis (TTS)
- ➢ Automatic Speech Recognition (ASR)

From information retrieval to thinking machines

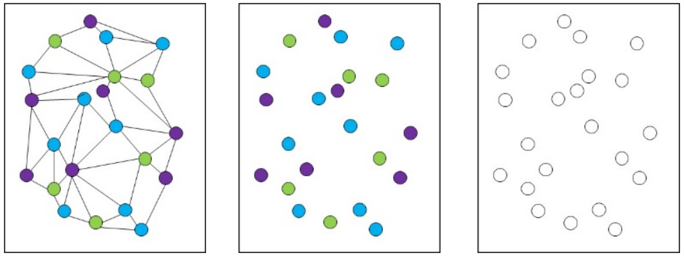
Combining speech with machine learning will lead to effective human-machine communication

Learn from human:
Data driven approaches

Human-like:
thinking machine

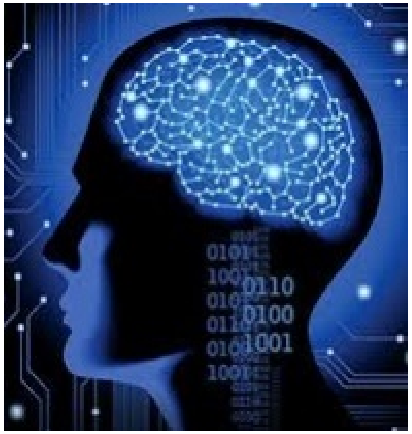


Knowledge   Ontology   Data



οντολογία: study categories of being and their relations

Numerical, Text, Spoken, …



1. natural, speech enabled, human-machine interface for information retrieval

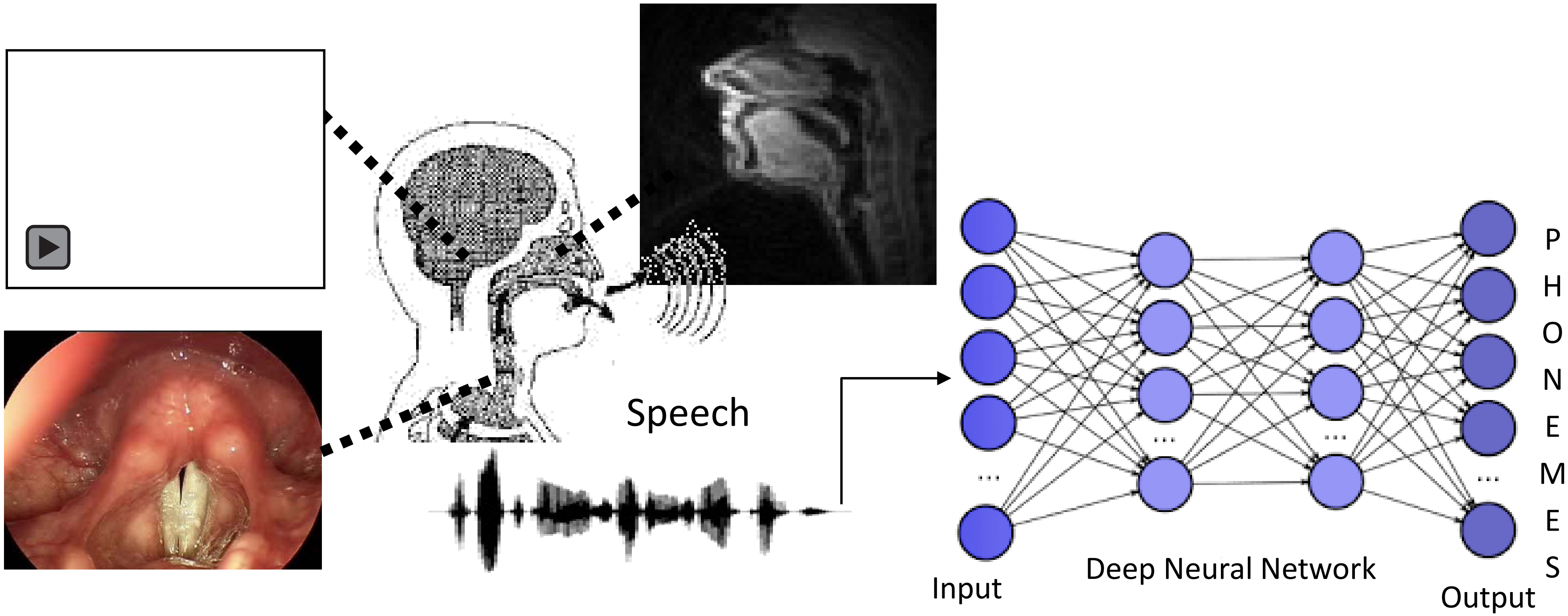
2. learn human's procedures

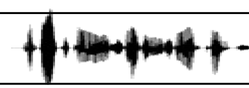
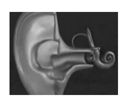
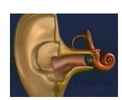
- make suggestions, compare, planning

❖ Design human centric information processing algorithms and services to create and access knowledge effectively, for improving productivity and quality of life

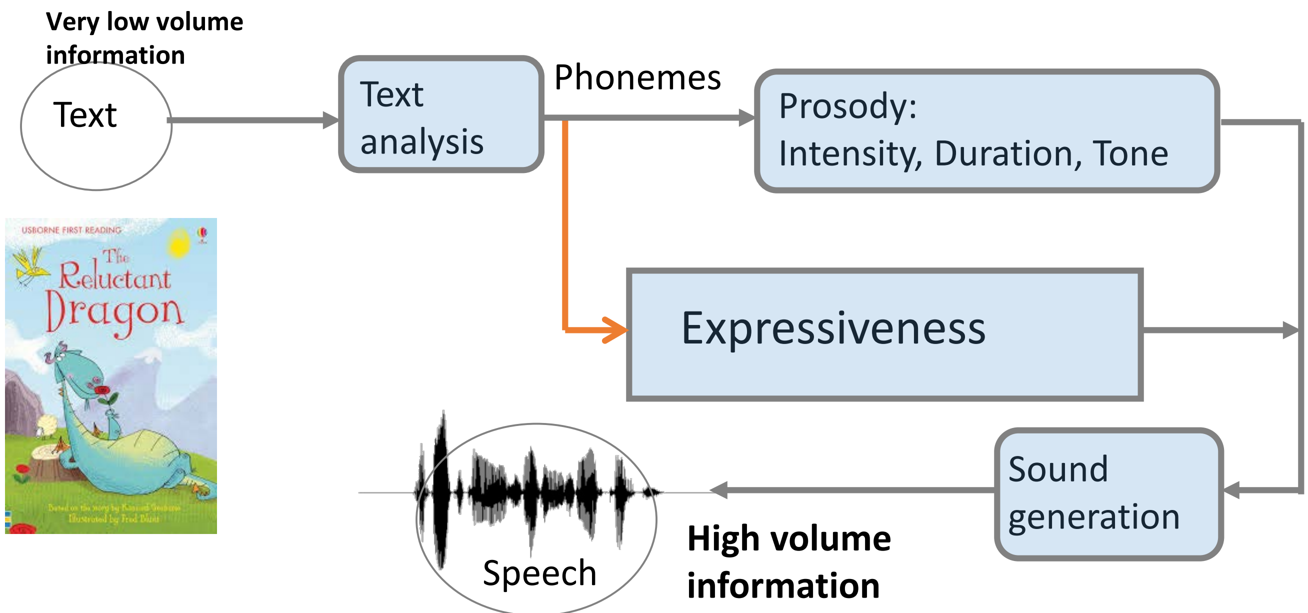
ASR: Automatic Speech Recognition; **NLU:** Natural Language Understanding;
NLG: Natural Language Generation; **TTS:** Text-to-speech

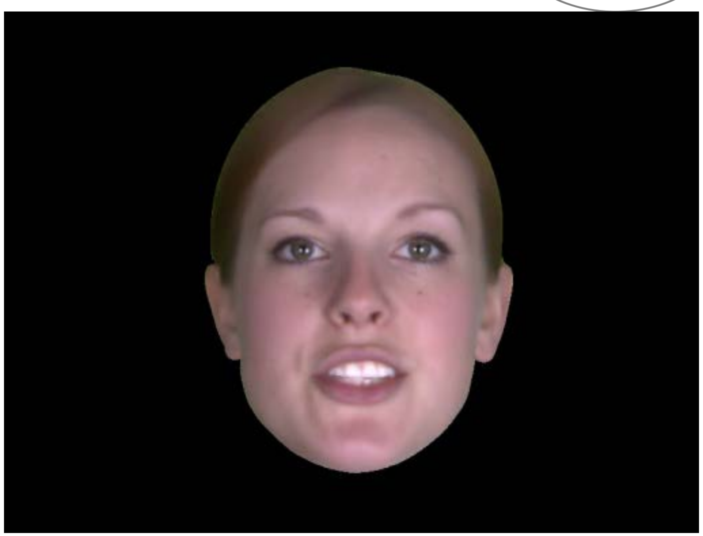
Automatic speech recognition: speech to text



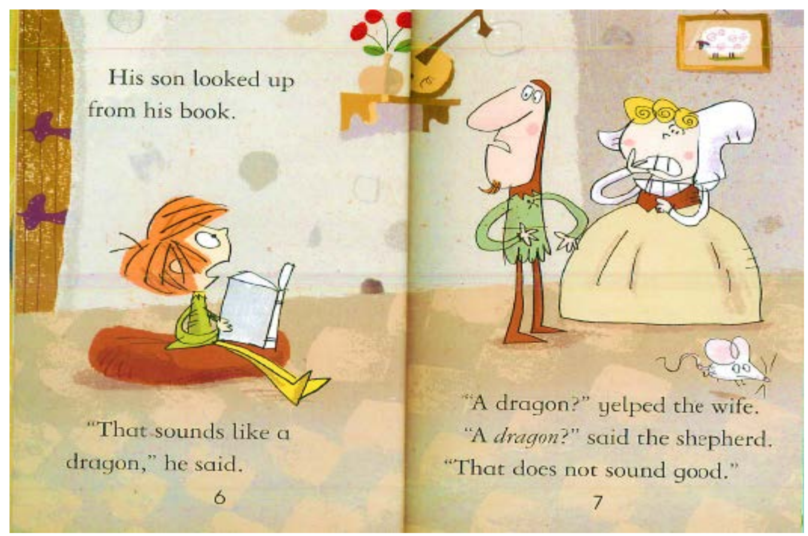
Approach		WER(%)	WER(%)
Waveform		42	58
State of art	 	14	34
CRL	  	12	33

Flexible and high quality visual text-to-speech synthesis



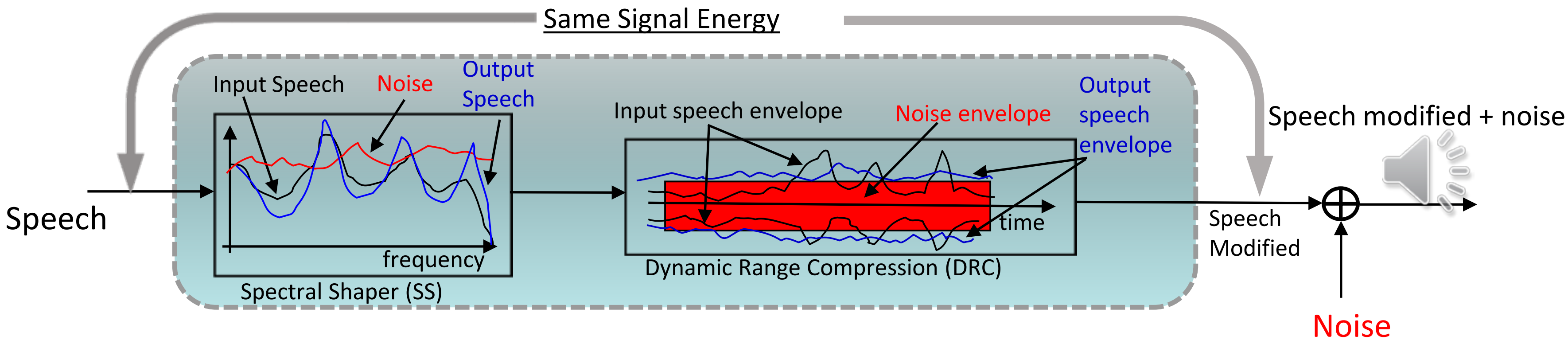


Xpressive Talk™



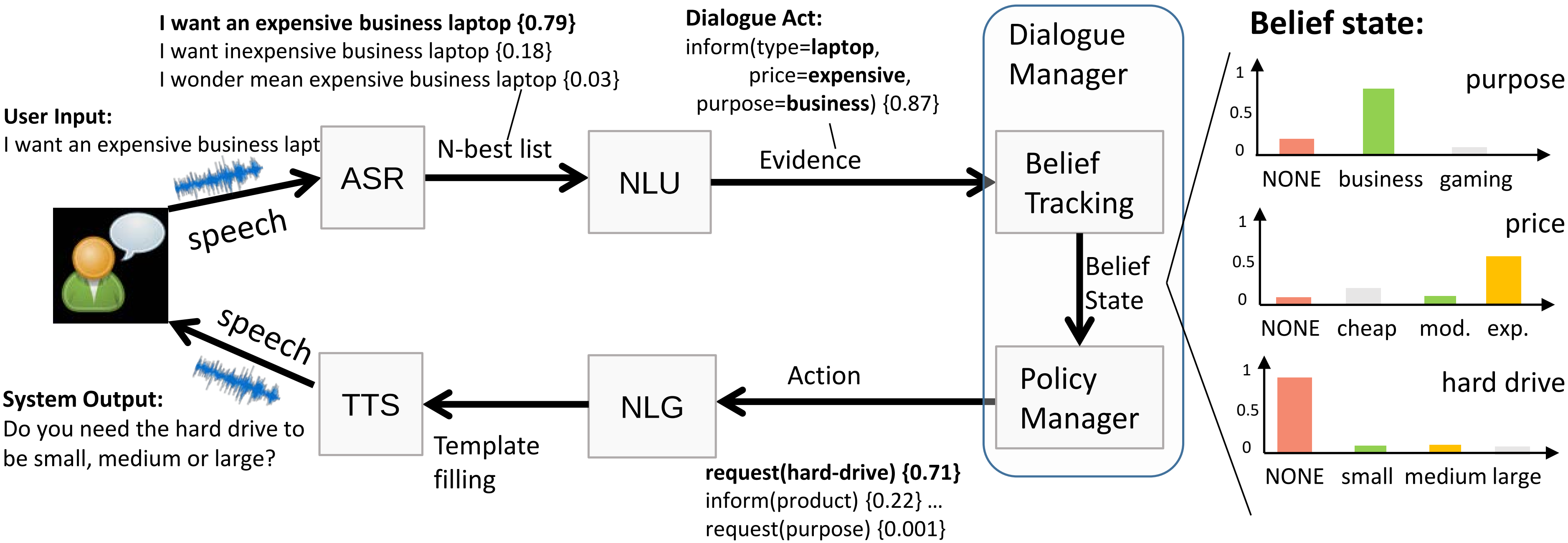
Intelligibility of speech in noise

- **Problem:** Speech Perception in Noise
- **Solution:** Spectral Shaping and Dynamic Range Compression (SSDRC)

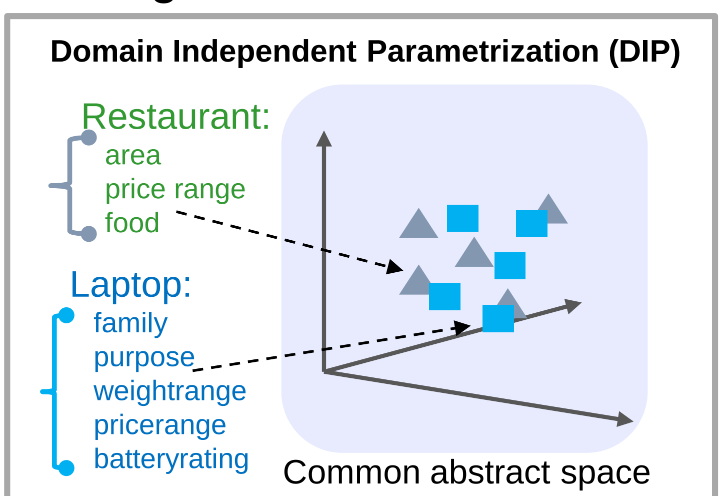


- **Applications:**

Statistical Dialogue Manager



➢ Transfer learning



	In Domain	Transfer learning
Success rate	85%	82%

ASR: Automatic Speech Recognition; **NLU:** Natural Language Understanding; **NLG:** Natural Language Generation; **TTS:** Text-to-speech

Example of natural human-machine communication

… with the CRL statistical spoken dialogue manager

Speech Signal Processing Lab

- a short tour

https://www.csd.uoc.gr/~sspl/index.html

Topics:

1. General overview, about us
2. Introduction to Speech Technology
3. Text to Speech Synthesis

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!

CONTEXT

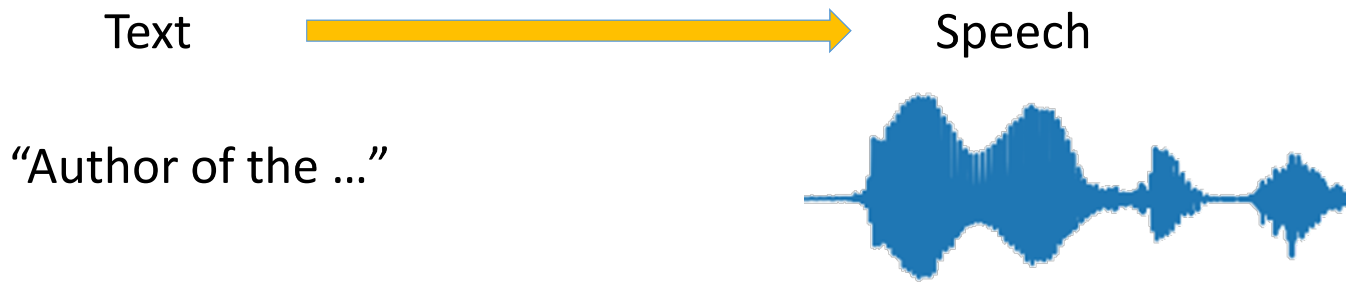
Text-to-Speech – the path so far

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

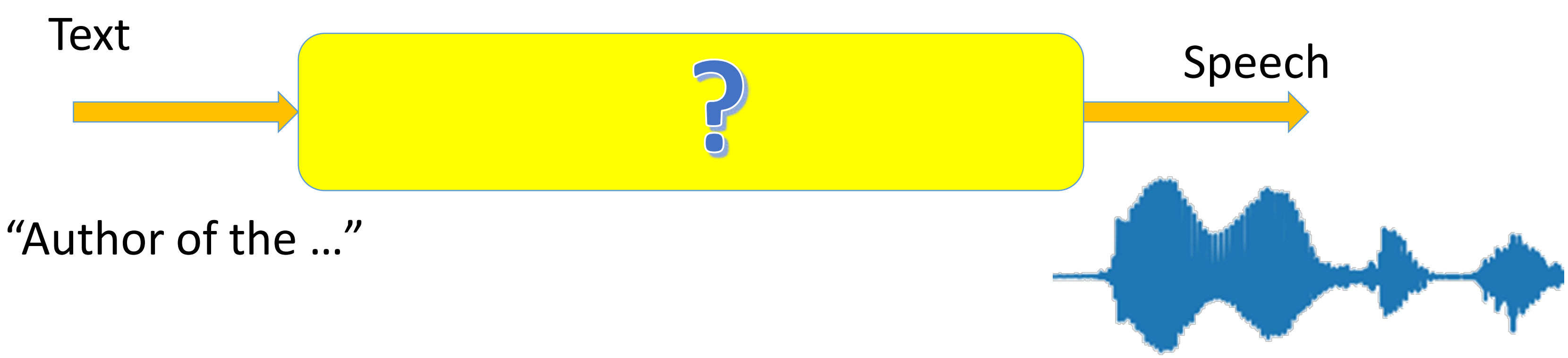
The first 3 audio files are from https://www.ims.uni-stuttgart.de/institut/mitarbeiter/moehler/synthspeech/#english

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

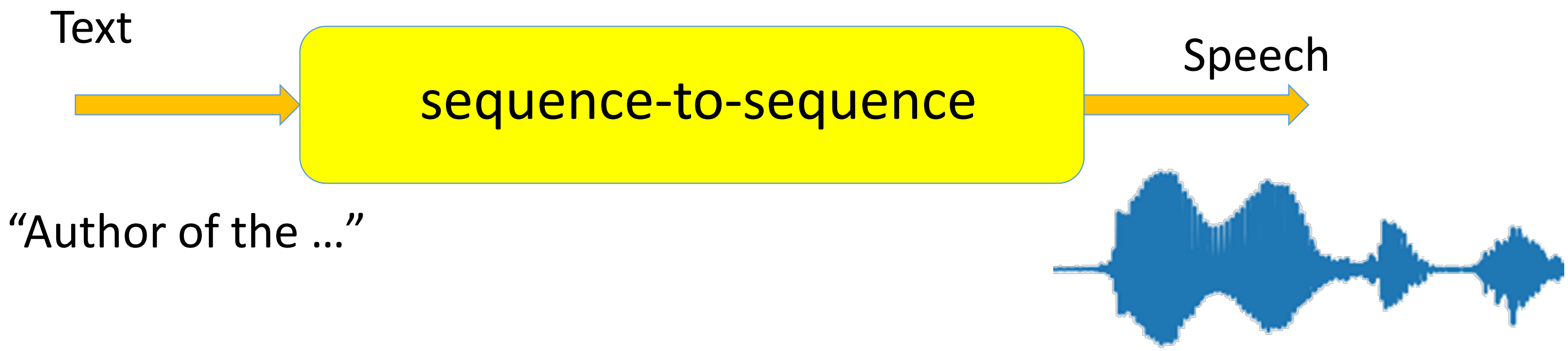
Text-to-Speech (as simple as that)



- End-to-end speech synthesis



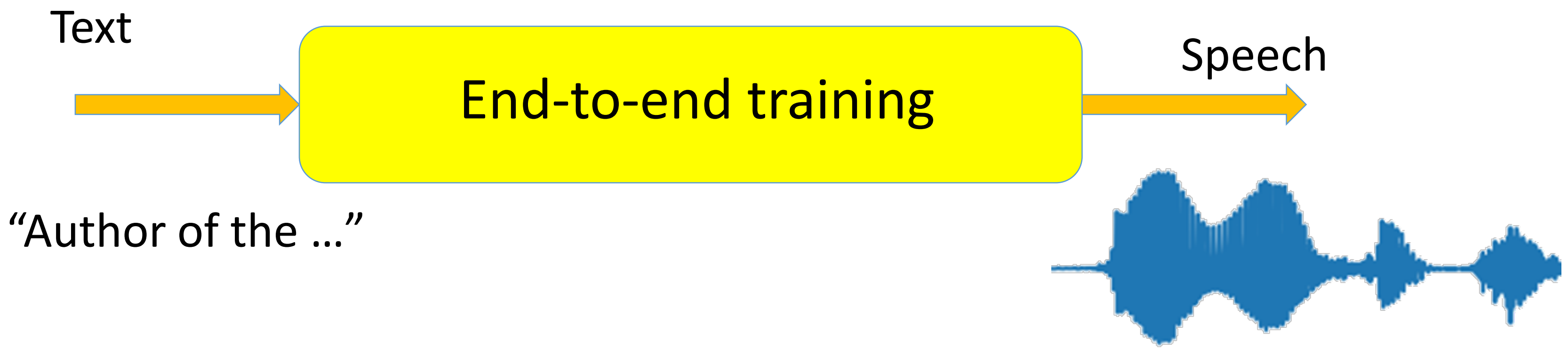
Text-to-Speech … a mapping problem

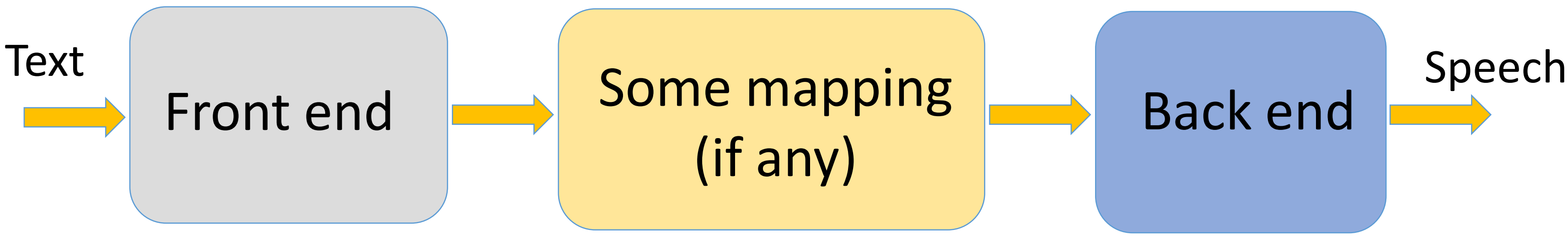


- ❑ Options:
  - Characters to samples
  - Phonemes to speech features and then to samples
  - Linguistic features to speech features and then to samples
  - Linguistic features to samples

A sequence to sequence problem

Text-to-Speech: the general framework





Outline

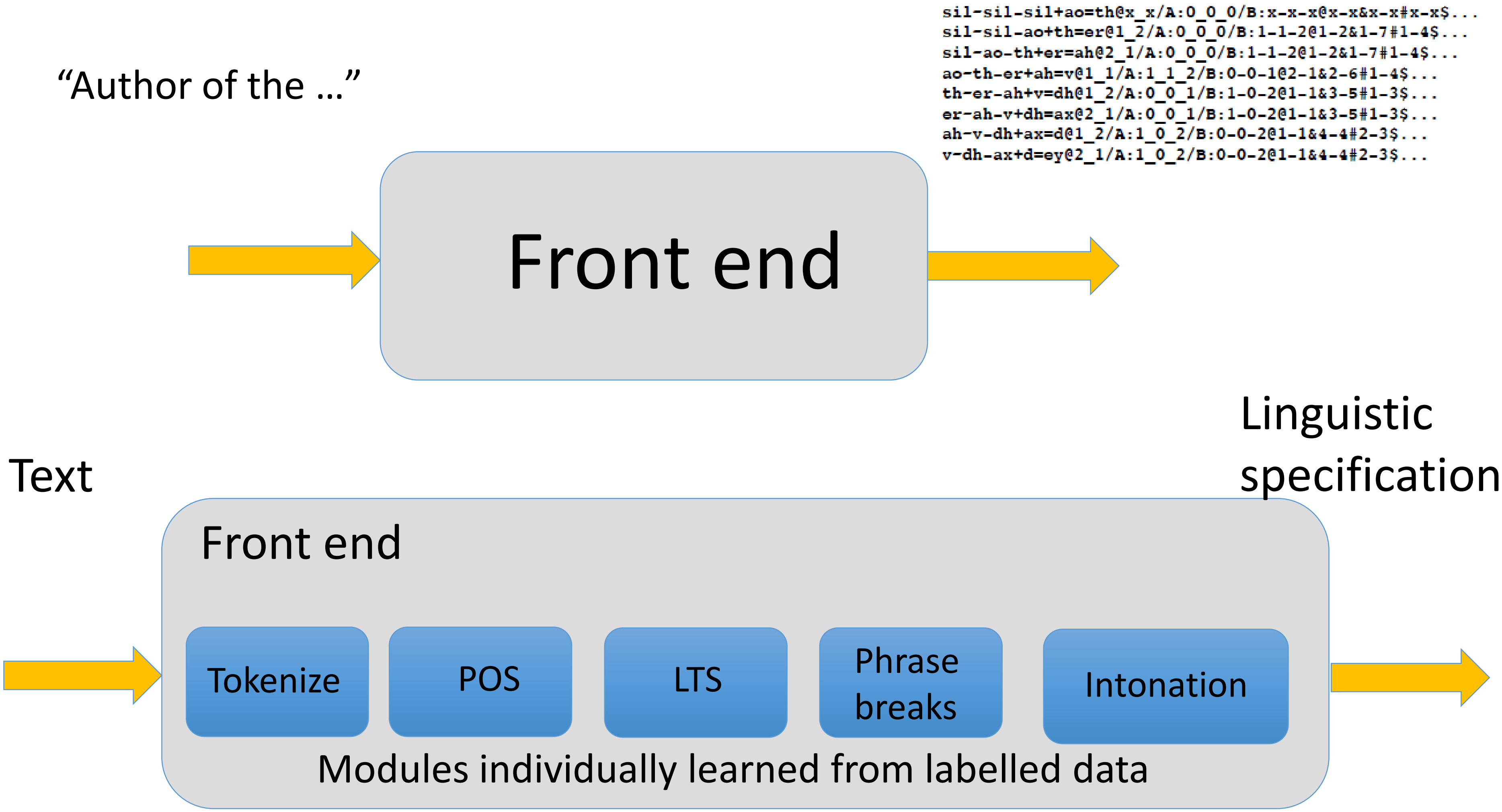
- Short overview
- Current concatenative systems – in a nutshell
- Statistical models – Regression
- Quick review of DNNs – a fast tour
- Neural TTS – sequence-to-sequence models
- Current Issues
- Applications
- Learning more …

Features from text - linguistics

"Author of the ..."

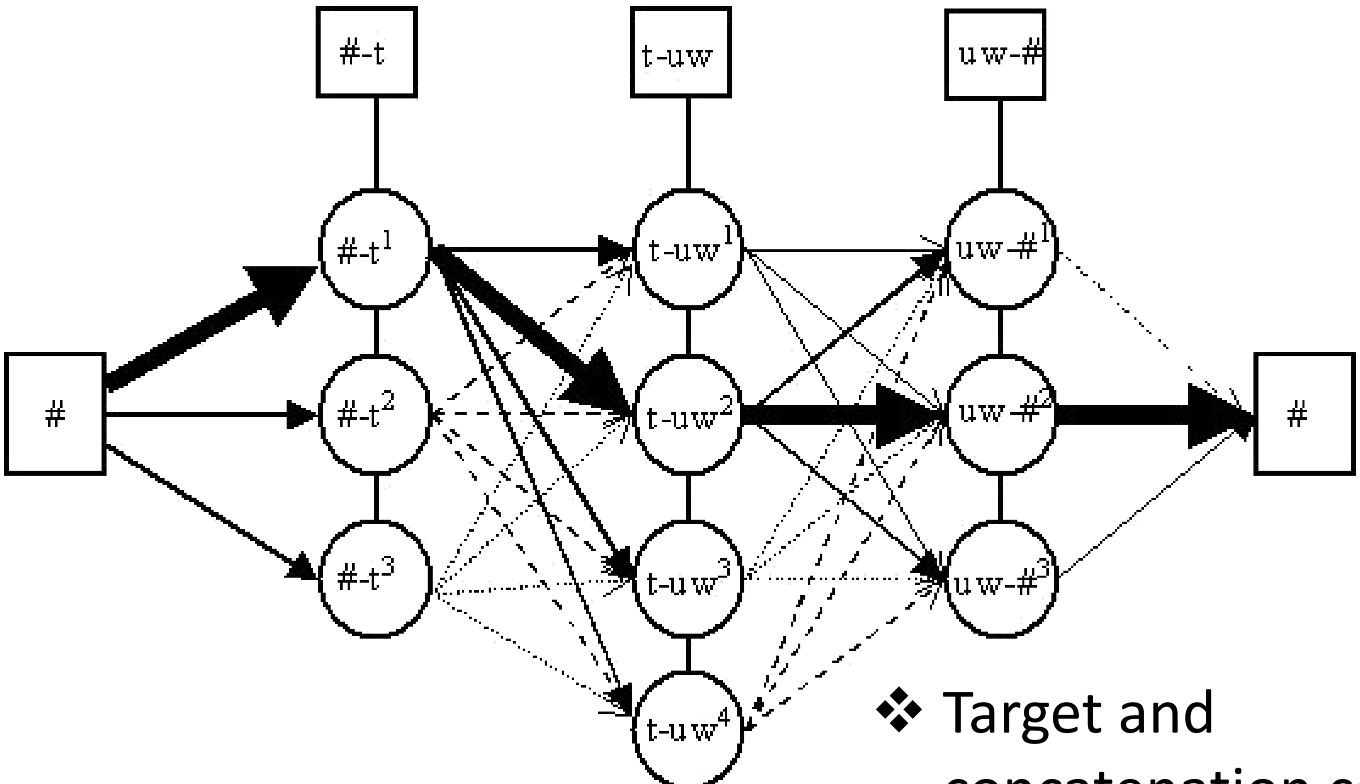
```
sil-sil-sil+ao=th@x_x/A:0_0_0/B:x-x-x@x-x&x-x#x-x$...
sil-sil-ao+th=er@1_2/A:0_0_0/B:1-1-2@1-2&1-7#1-4$...
sil-ao-th+er=ah@2_1/A:0_0_0/B:1-1-2@1-2&1-7#1-4$...
ao-th-er+ah=v@1_1/A:1_1_2/B:0-0-1@2-1&2-6#1-4$...
th-er-ah+v=dh@1_2/A:0_0_1/B:1-0-2@1-1&3-5#1-3$...
er-ah-v+dh=ax@2_1/A:0_0_1/B:1-0-2@1-1&3-5#1-3$...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



Concatenative systems (pure)

- From linguistic features to units (samples)

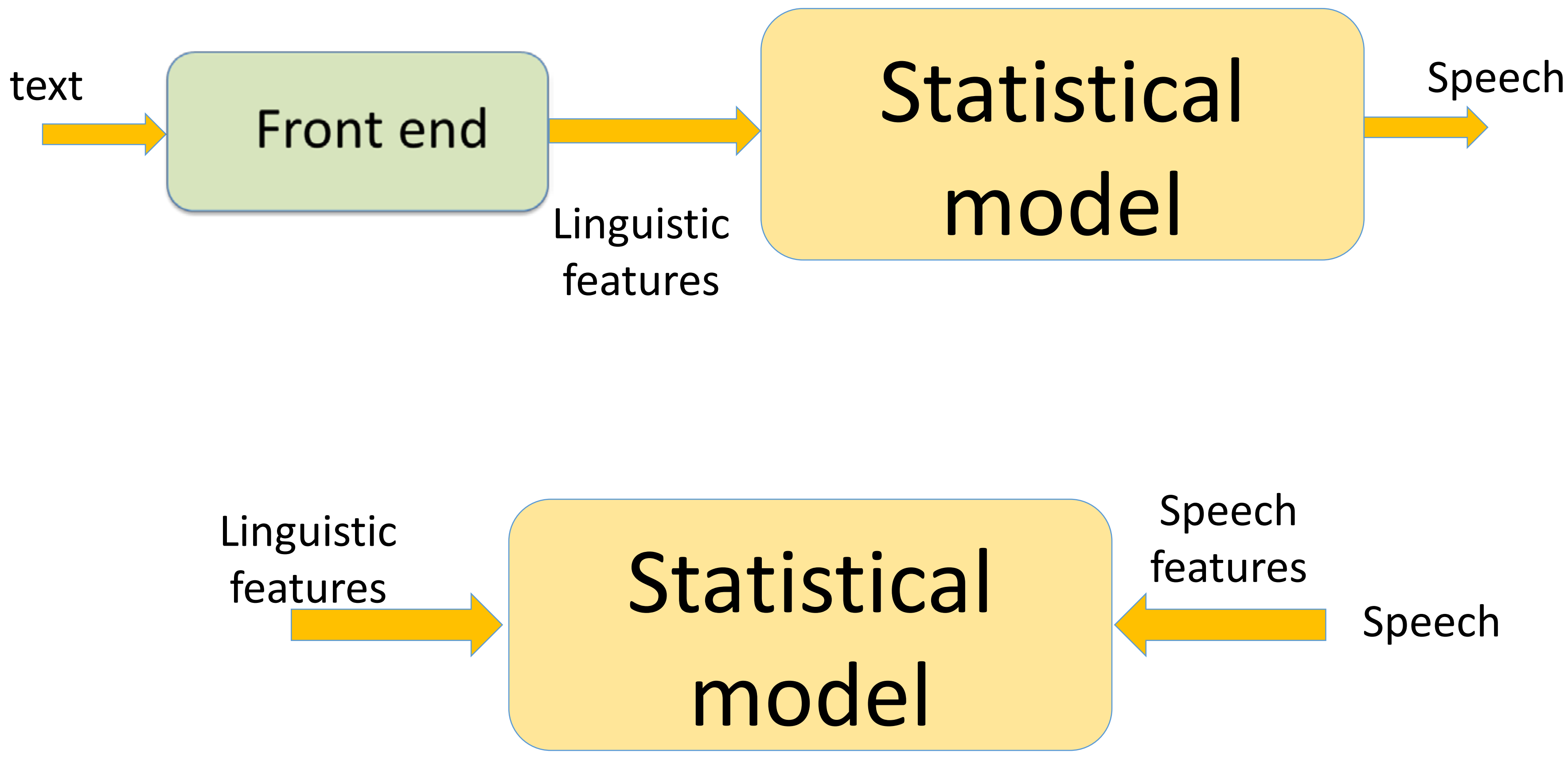


- Target and concatenation costs

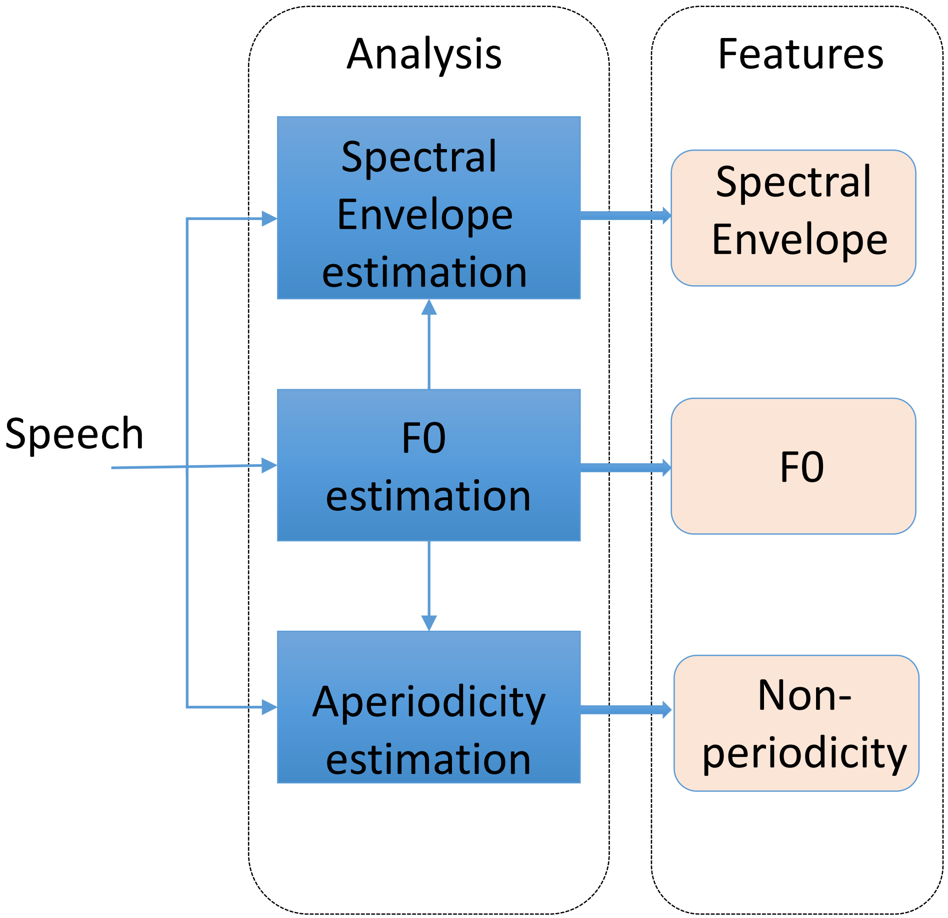
Outline

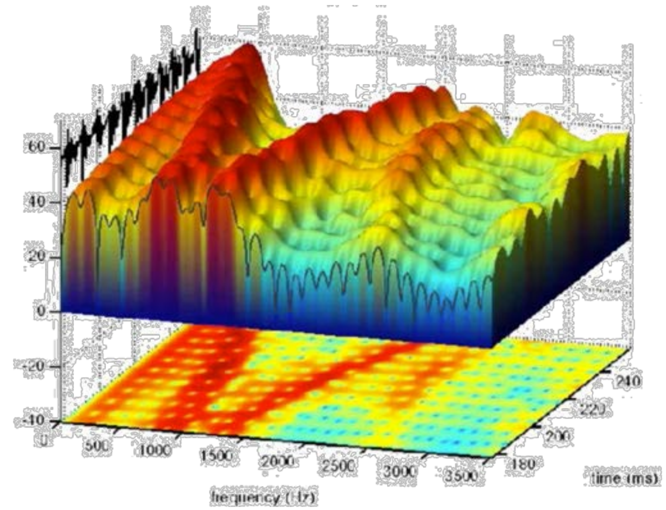
- Short overview
- Current concatenative systems – in a nutshell
- Statistical models – Regression
- Quick review of DNNs – a fast tour
- Neural TTS – sequence-to-sequence models
- Current Issues
- Applications
- Learning more …

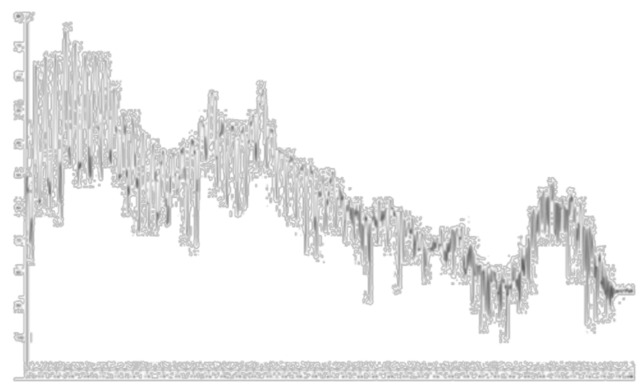
Start learning from data

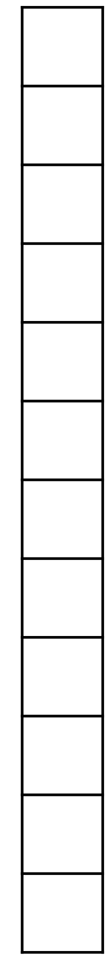


Speech features – STRAIGHT (H. Kawahara)









Start learning from data

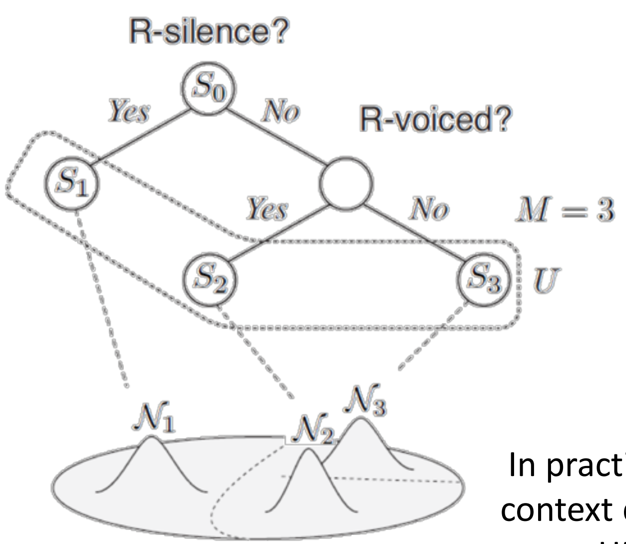
Linguistic features

Statistical model

Speech features

Speech

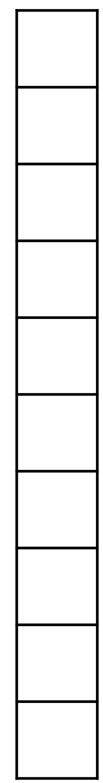
Decision Tree



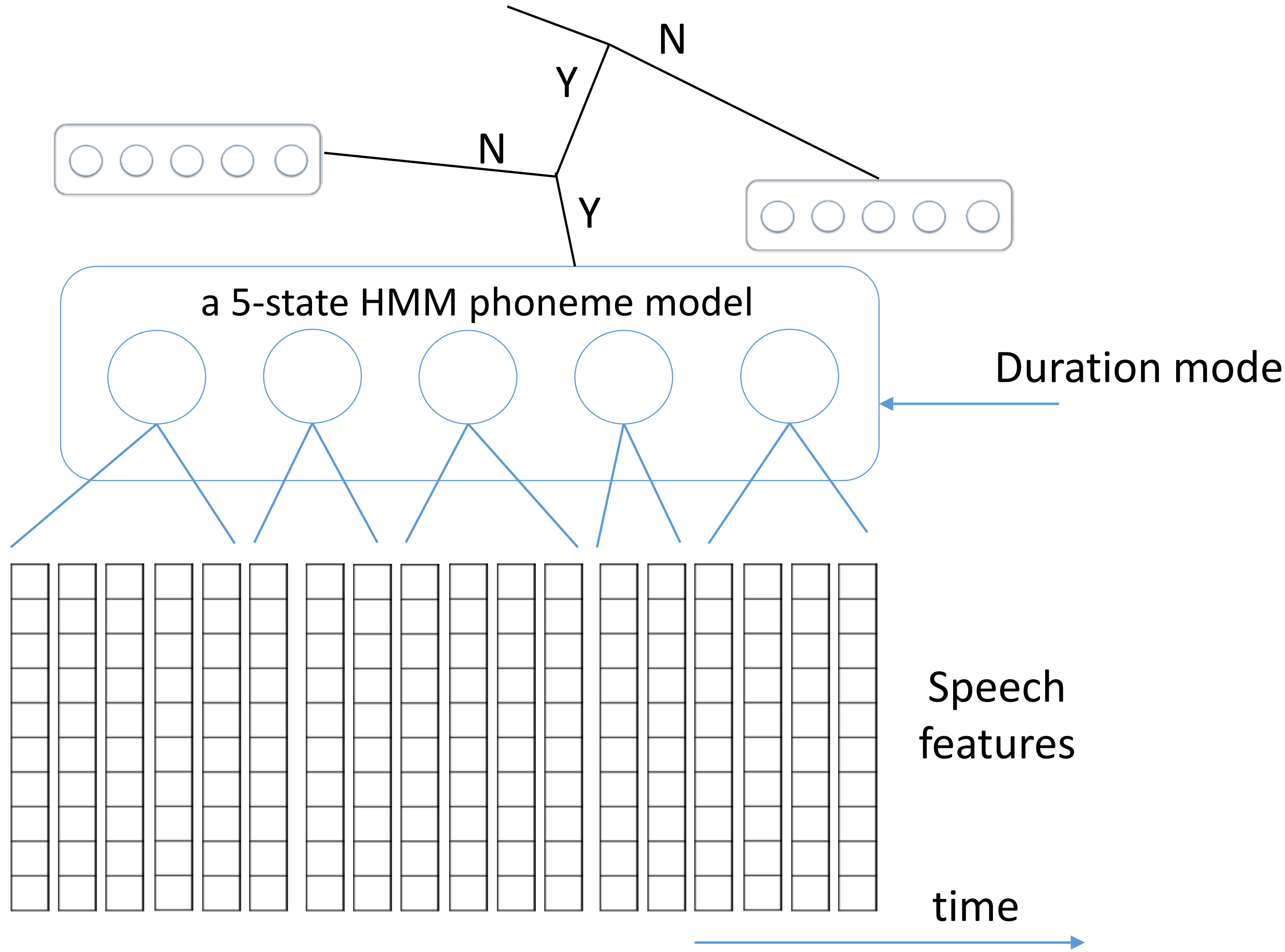
Regression

In practice we use context dependent HMMs

Figure from J. Yamagishi



Text-to-features using CART

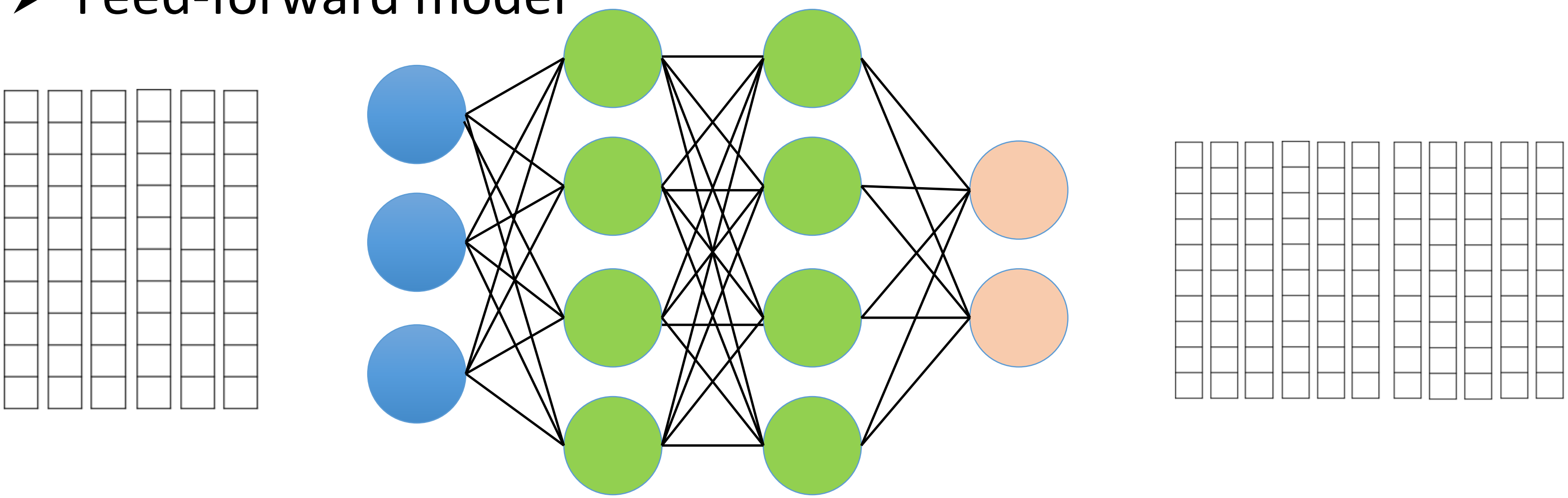


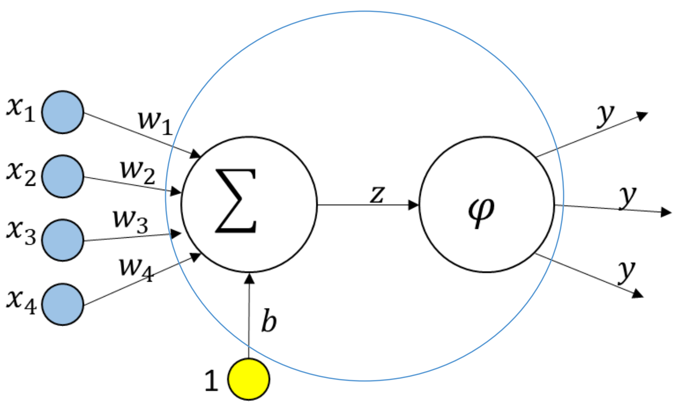
Outline

- 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

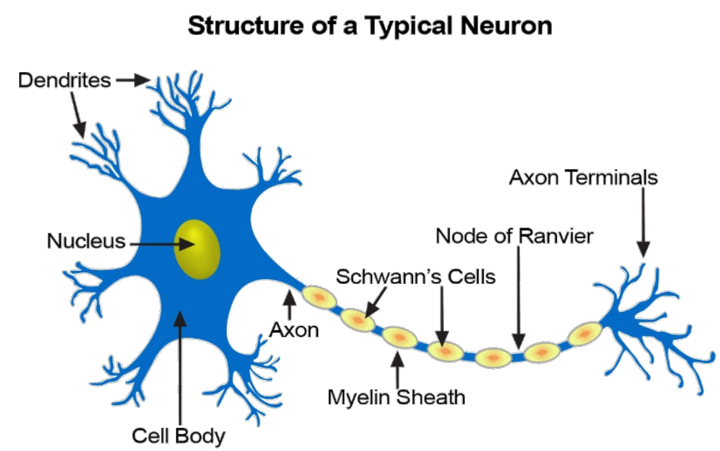
- Feed-forward model





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

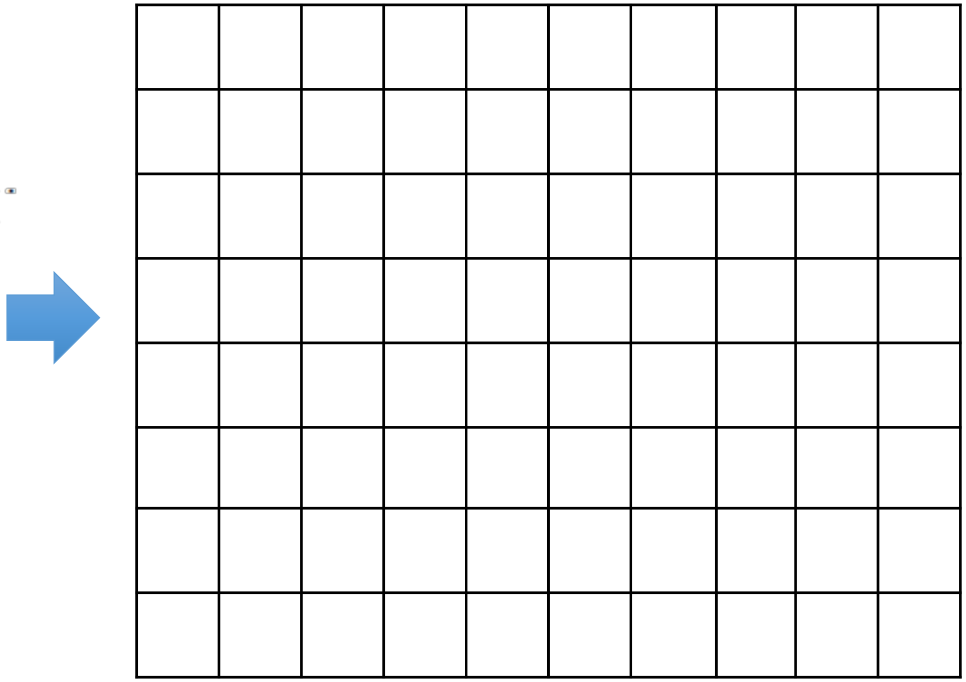
$$y = \varphi(z) = \begin{cases} 0, & z < 0 \\ 1, & z \geq 0 \end{cases}$$



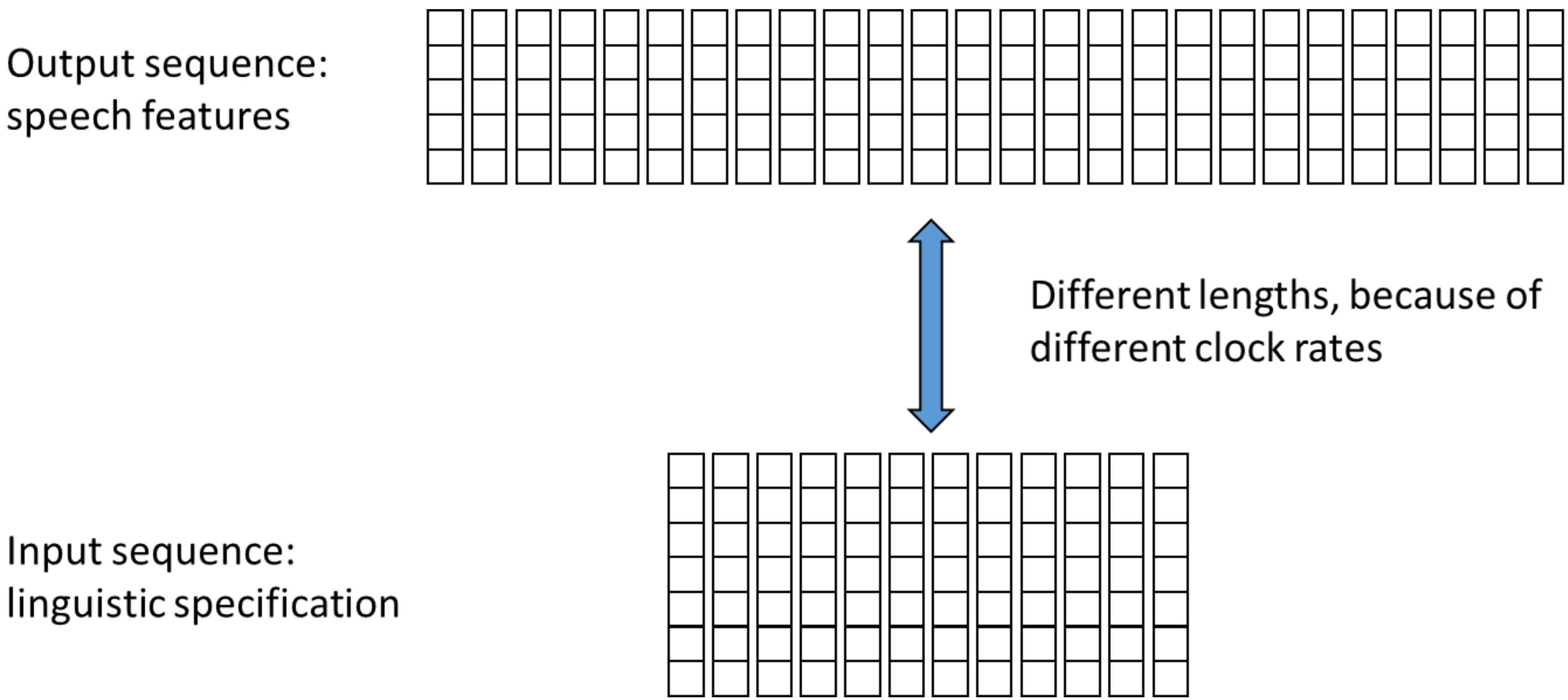
Going back to our problem: TTS (with DNNs)

- 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$...
v-dh-ax+d=ey@2_1/A:1_0_2/B:0-0-2@1-1&4-4#2-3$...
```

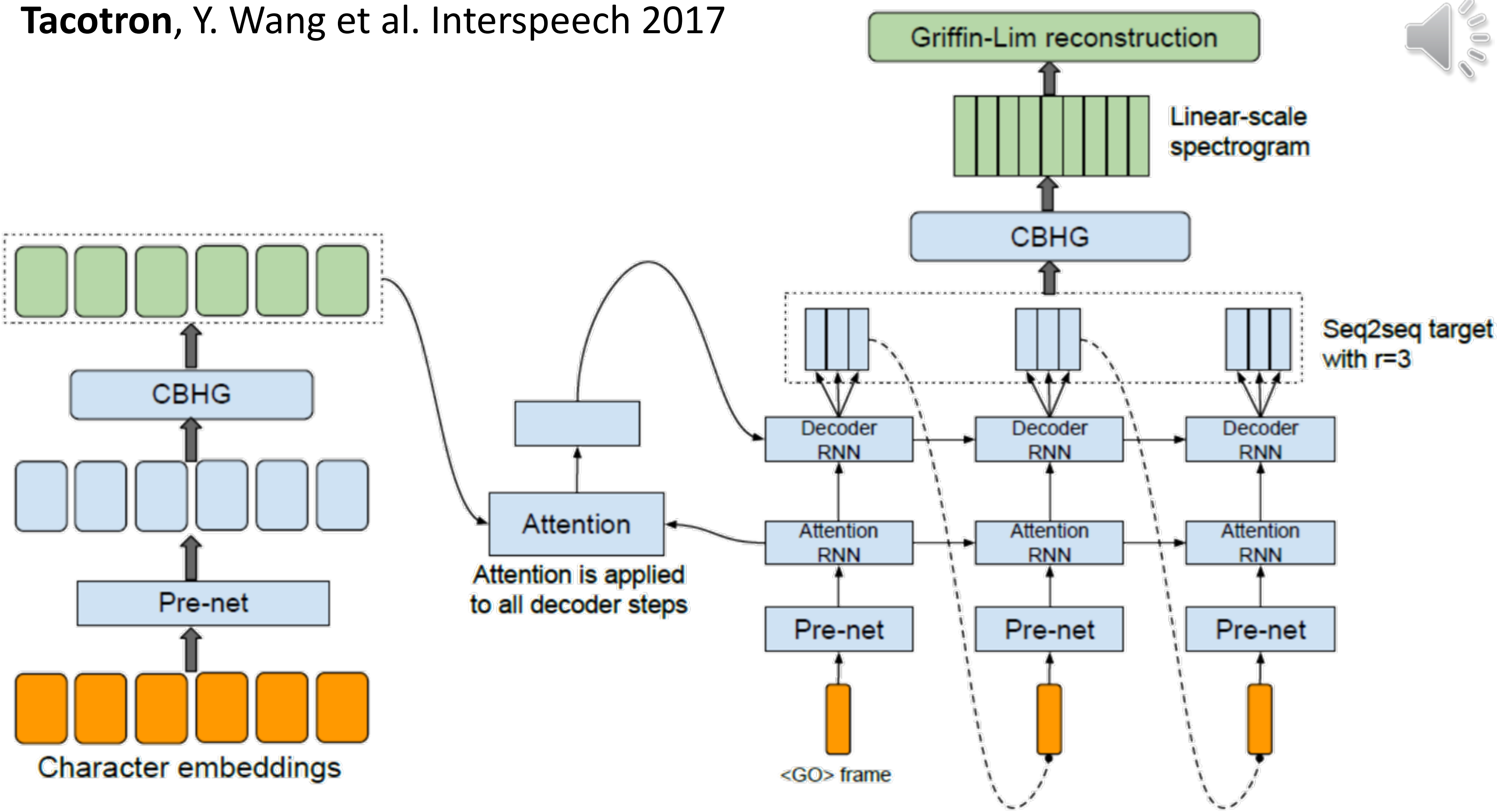


Neural TTS = a sequence-to-sequence regression



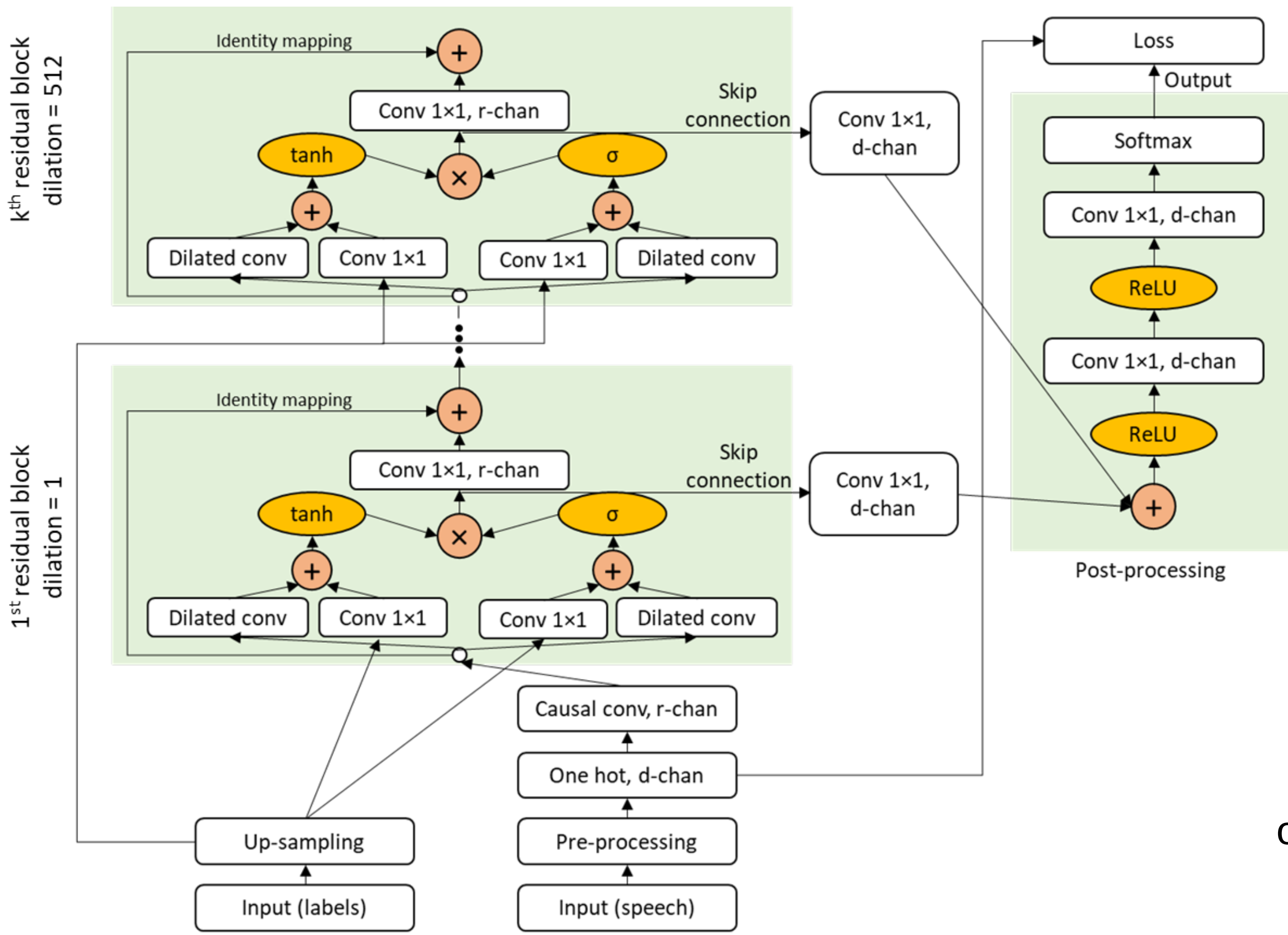
Tacotron: a multiple sequence-to-sequence model

Tacotron, Y. Wang et al. Interspeech 2017



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

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



Sound examples
(16 kHz) [test data]:

- ❖ with natural prosody:
  - o Google (40 hours)
  - o UoC (5 hours)
- ❖ with synthetic prosody (HMM):
  - o UoC (5 hours)

Sound examples from Univ. of Crete trained on vocoded speech

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 …

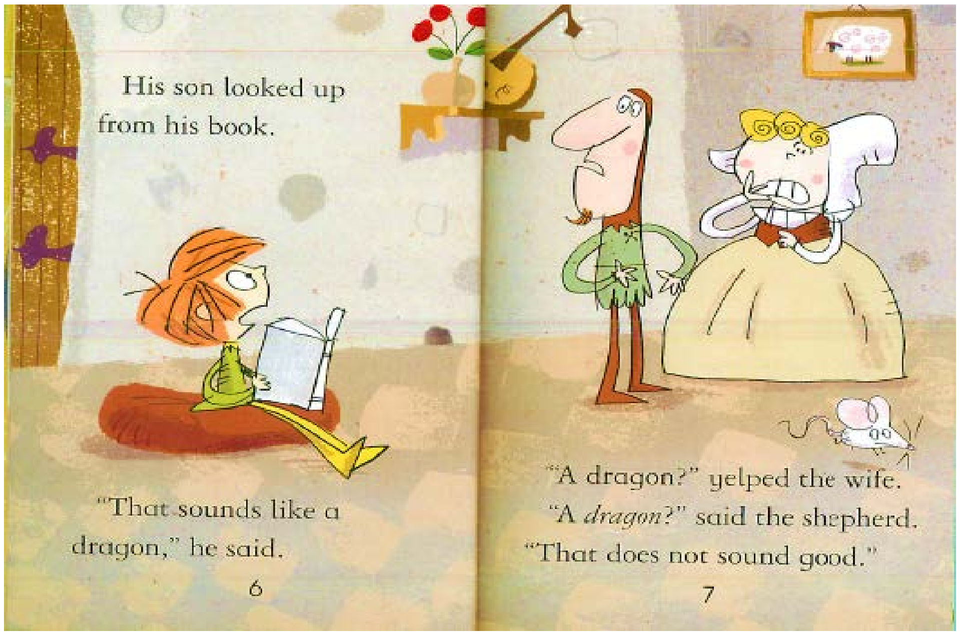
Speech Synthesis – current issues

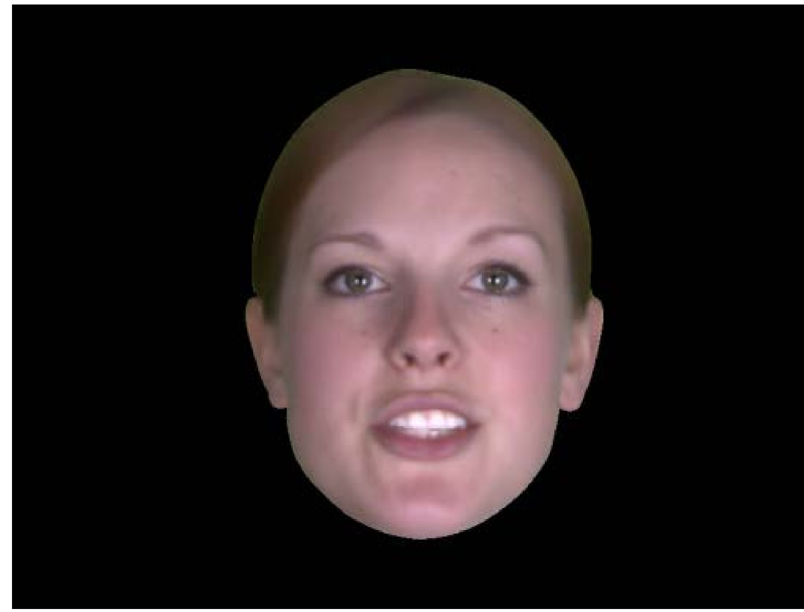
- Robustness & running cost
  - Robust & fast front-end and back-end (Parallel Wavenet, WaveRNN, …)
  - Robust to recordings quality and quantity
  - Robust training
- Context awareness
  - Adaptation to user acts in dialogue (conversational TTS, style token)
  - Adaptation to the listening conditions (intelligibility)

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.

BBC NBC sky NEWS REUTERS

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 …

Speech Processing Courses in Crete

SPCC

July 27-31, 2020

Crete, Greece

- Learn (with theory in the mornings and hands on in the afternoons) about:
  - ✓ Neural Source-Filter vocoders for synthesis (Junichi Yamagishi and Xin Wang, NII Japan)
  - ✓ Sample, autoregressive neural vocoders (Vassilis Tsiaras, UoC, Greece)
  - ✓ Neural Vocoders for coding (Jan Skoglund, Google, USA)
  - ✓ Neural based speech enhancemt (Paris Smaragdis, Univ of Illinois, USA)