ML for Audio Study Group

Session 3: TTS Deep Dive

Jan 4th, 2022, 5 PM CET

hf.co/join/discord

Vatsal Aggarwal

Vaibhav (VB)
Srivastav

JOIN

Suggested readings before this session

- Notebooks (link - https://github.com/Vaibhavs10/ml-with-audio)
  - Intro to Audio data notebook
  - Intro to ASR Notebook
- Speech and Language Processing 26.6

Introduction

Vatsal Aggarwal

(https://www.linkedin.com/in/vatsal-aggarwal-993472104/)

- DL based vocoding in production
- Zero-shot speech generation

Vaibhav Srivastav (https://twitter.com/reach_vb)

- MS student @ Uni Stuttgart/ Working Student @ Deloitte Tax
- Previously
  - Strategy @ Deloitte Consulting

Organisation

- **Community-led!**
  - We’ll kick off with some basics, but we’ll decide collaboratively where we want to focus
  - Anyone can participate!
  - Members of the HF team and other cool collaborators will join.
- Expectation
  - Before each session: **Read/watch related resources**
  - During each session, you can
    - Ask question in the forum
    - Present a short (~10-15mins) presentation on the topic (agree beforehand)
    - Participate a bit more passively (that’s also ok and you’re welcomed!)
  - Before/after:
    - Keep discussing/asking questions about the topic
    - Share interesting resources

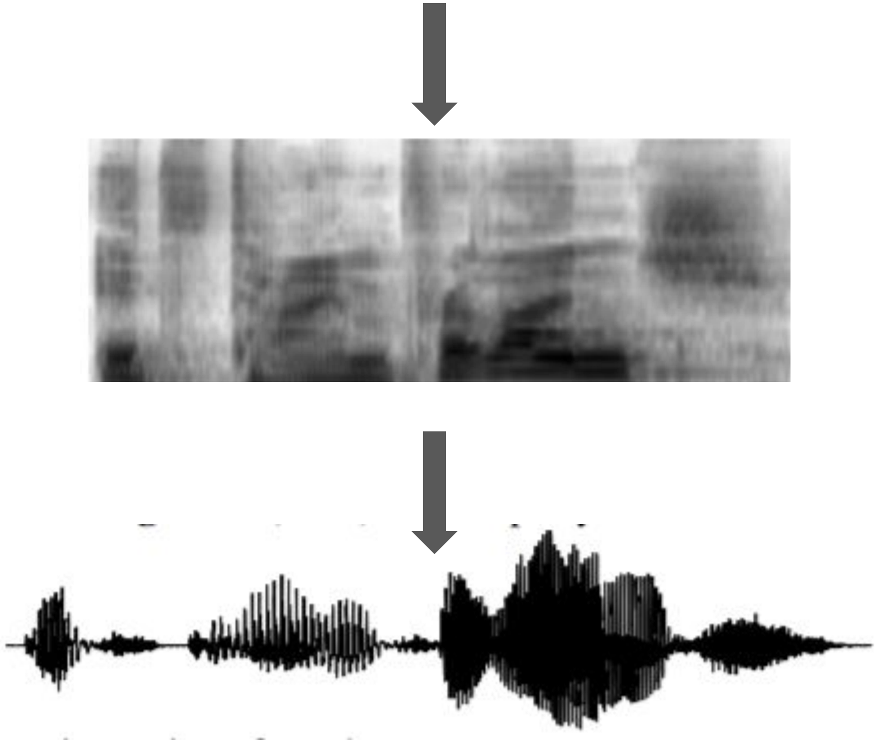
Timeline

- Dec 14: Kick off session
- Dec 21: ASR Deep Dive
- **Jan 4: TTS Deep Dive**
- Jan 18: pyctcdecode: A simple and fast speech-to-text prediction decoding algorithm

Text-to-Speech

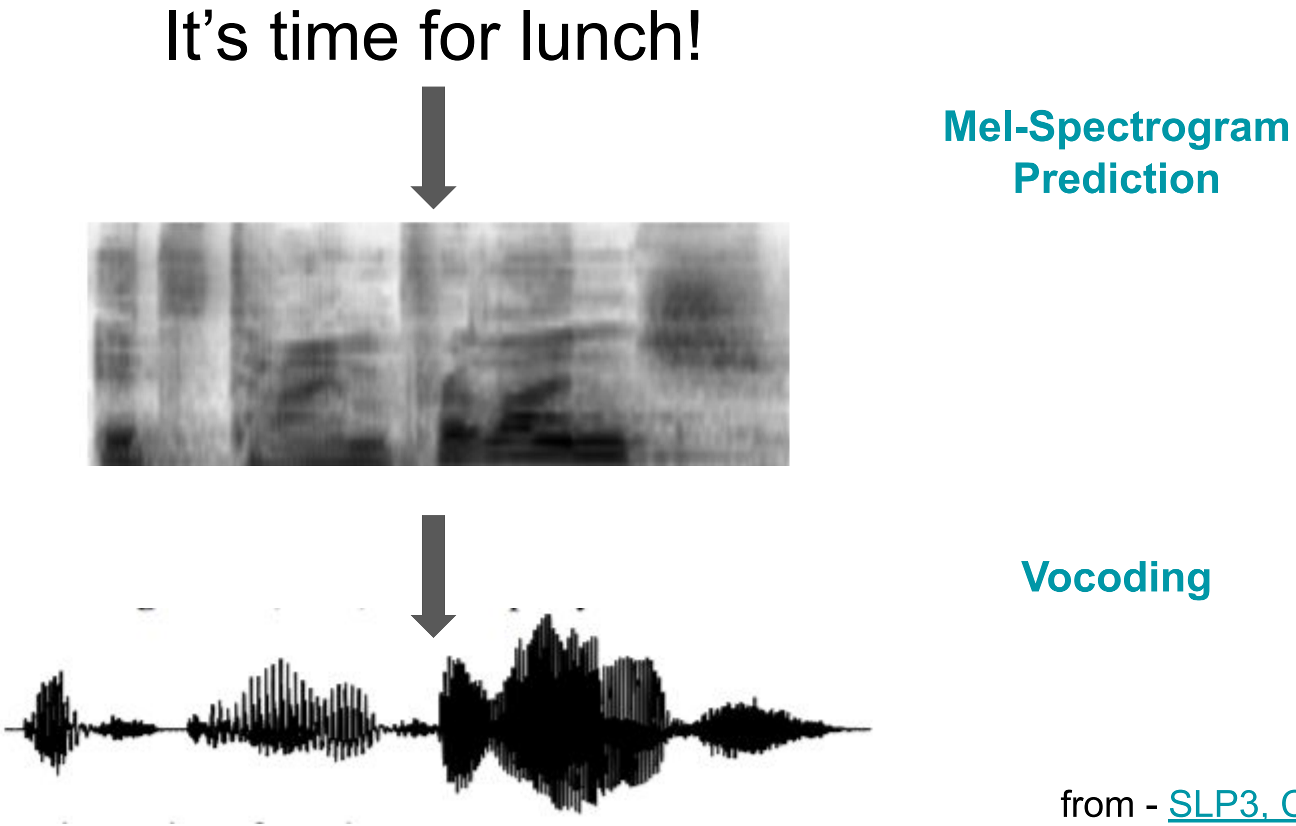
Text to Speech

It’s time for lunch!



from - SLP3, Ch 26

Text to Speech



It’s time for lunch!

Mel-Spectrogram Prediction

Vocoding

from - SLP3, Ch 26

TTS | Why is TTS hard?

It’s no **use** to ask to **use** the telephone.

Do you **live** near a zoo with **live** animals.

I prefer **bass** fishing to playing the **bass** guitar.

TTS | Why is TTS hard?

It’s no **use (/y uw s/)** to ask to **use (/y uw z/)** the telephone.

Do you **live (/l ih v/)** near a zoo with **live (/l ay v/)** animals.

I prefer **bass (/b ae s/)** fishing to playing the **bass (/b ey s/)** guitar.

TTS | Text Normalisation

TTS systems require preprocessing for handling non-standard words:

1. numbers
2. monetary amounts
3. abbreviations
4. dates
5. acronyms, etc

TTS | Text Normalisation

TTS systems require preprocessing for handling non-standard words:

seventeen fifty: (in “The European economy in 1750”)

one seven five zero: (in “The password is 1750”)

seventeen hundred and fifty: (in “1750 dollars”)

one thousand, seven hundred, and fifty: (in “1750 dollars”)

TTS | How exactly is this solved?

Modern end-to-end TTS systems can learn to do some normalization themselves however, due to limited amount of training data, a separate normalization step is needed.

1. **Rules** (ex: regex)
2. **Seq2Seq** model (requires a bit more post processing)

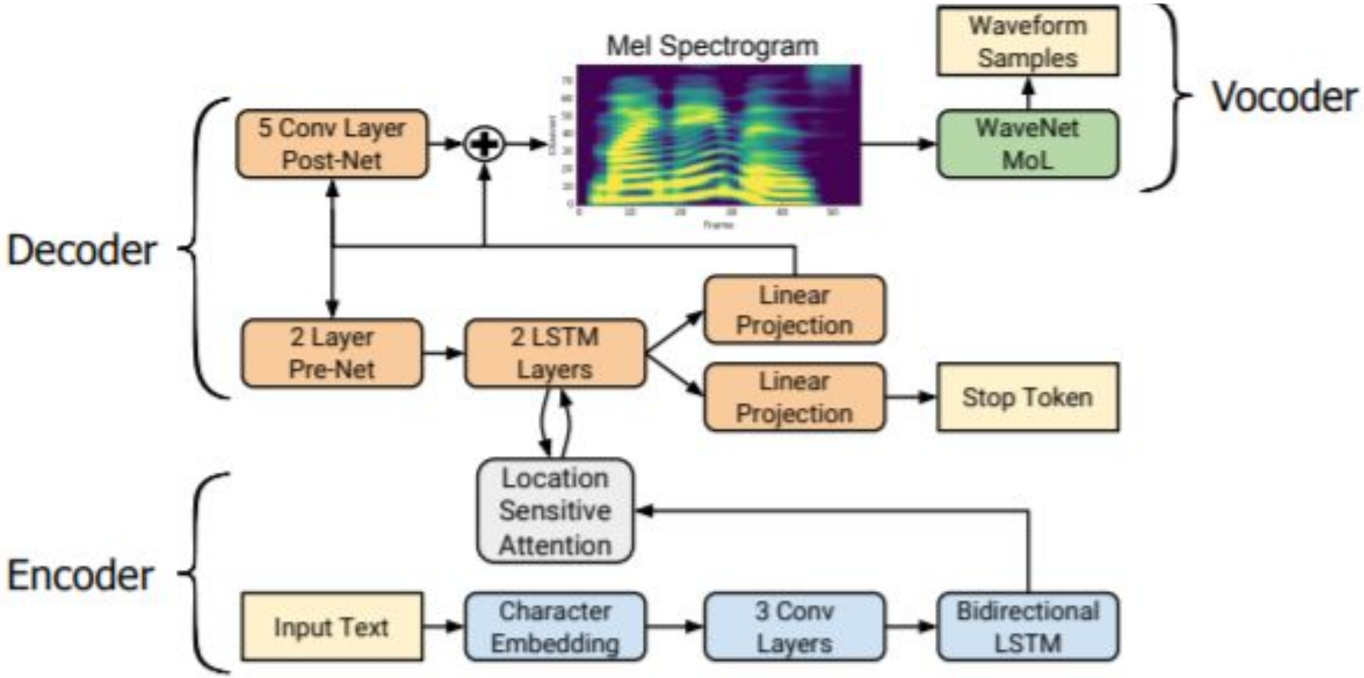
from - SLP3, Ch 26

TTS | Mel-Spectrogram Prediction

1. same architecture as ASR - *encoder-decoder with attention*
2. the encoder takes a sequence of letters and produce a hidden representation representing the letter sequence
3. the hidden representation is then used by the attention mechanism in the decoder

from - SLP3, Ch 26

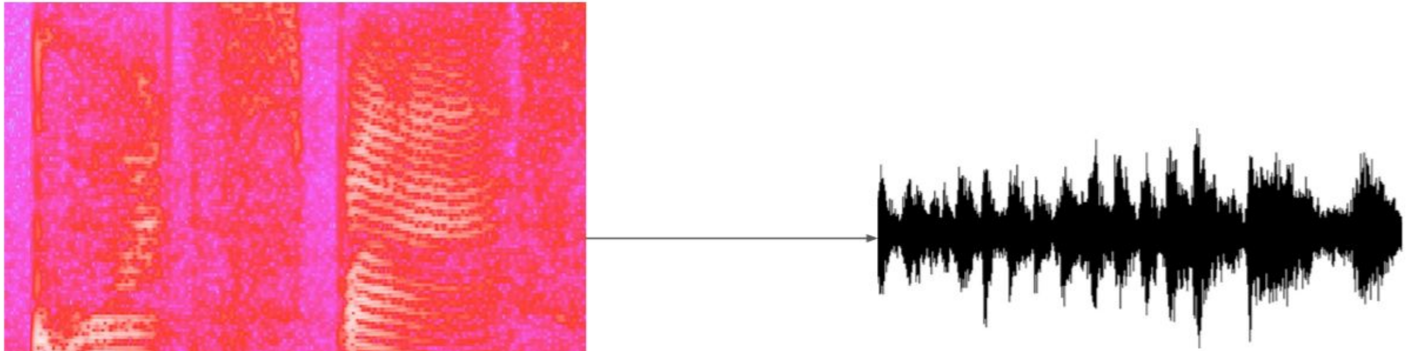
TTS | Tacotron 2



from - SLP3, Ch 26

TTS | Vocoding

Goal: to invert a log mel spectrum representations back into a time-domain waveform representation



from - SLP3, Ch 26

TTS | Vocoding

Cue.. **Wavenet**

- takes spectrograms as input and produces sequences of 8-bit mu-law (audio)
- many layers of dilated convolutions for a high receptive field
- output of the dilated convolutions is passed through a softmax which makes this 256-way decision (8-bit)

from - SLP3, Ch 26

TTS | Wavenet

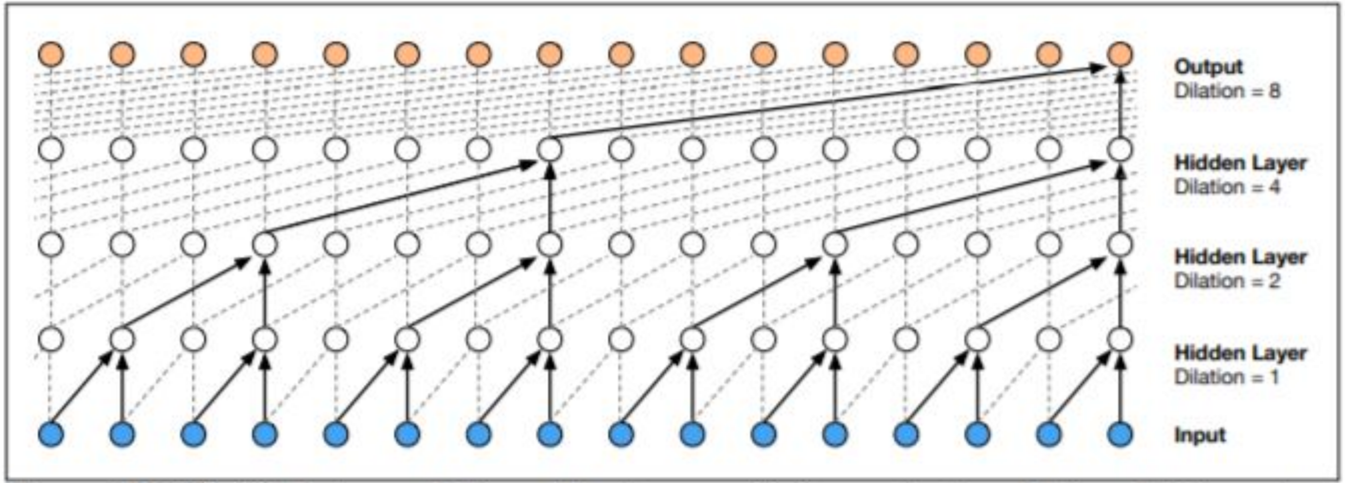


Figure 26.15 Dilated convolutions, showing one dilation cycle size of 4, i.e., dilation values of 1, 2, 4, 8. Figure from van den Oord et al. (2016).

from - SLP3, Ch 26

TTS | Evaluation

Mean Opinion Scores (MOS) - a rating of how good the synthesized utterances are, usually on a scale from 1–5.

AB Tests - play the same sentence synthesized by two different systems. The human listeners choose which of the two utterances they like better.

from - SLP3, Ch 26

TTS | Model Types

Text to Spectrogram Models

- Attention-based (*e.g. Tacotron*)
- Duration-based (*e.g. FastSpeech*)

Spectrogram to Waveform Models

- Autoregressive (*e.g. WaveNet, WaveRNN*)
- Flows (*e.g. WaveGlow, Parallel WaveNet*)
- GANs (*e.g. MelGAN, Parallel WaveGAN*)

Intro

@vatsal_aggarwal

feedback.vatsal.io

History/Journey of Speech Synthesis

“Aim” of Speech Research

- Naturalness
- Intelligibility
- Prosody/Expressivity
- Amount of data required
- Adaptation to situation
- Ethical use

Deep Learning made significant progress in producing more “natural” synthetic speech whilst enabling better flexibility (e.g. expressivity) and lower amounts of data.

Fun fact: “Vocoder”

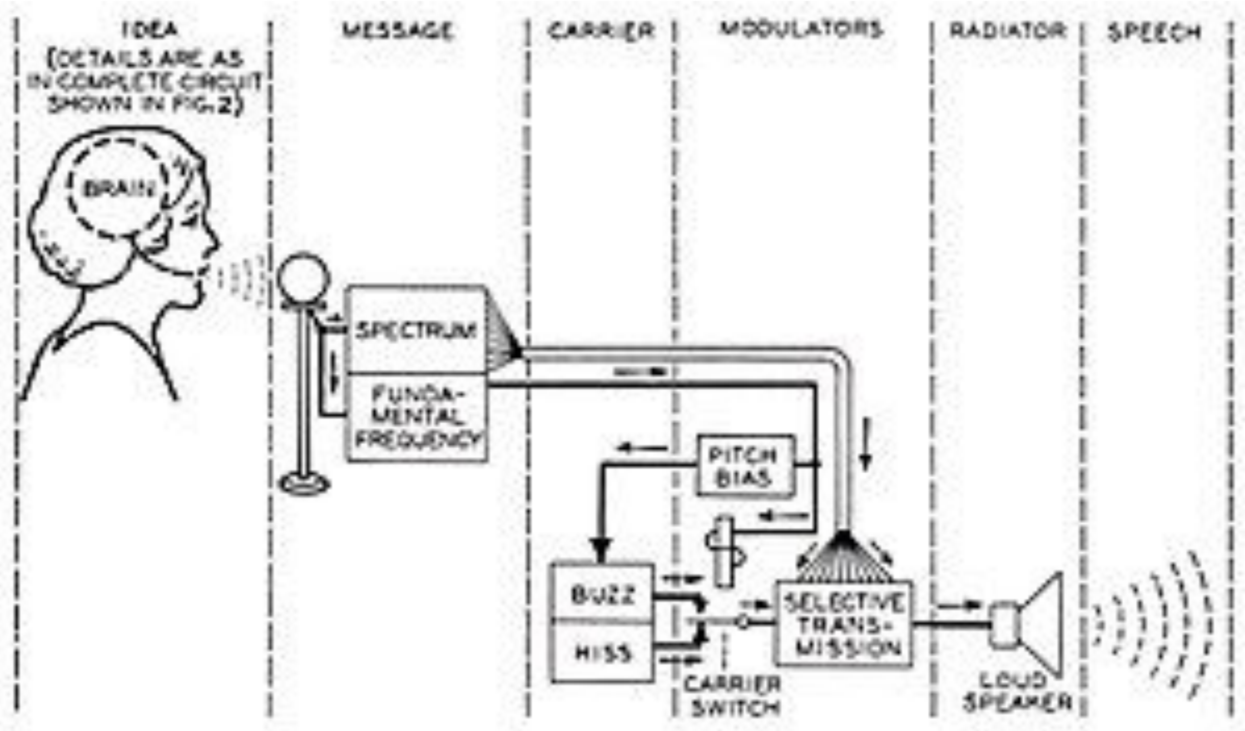


Fig. 7—Schematic circuit of the vocoder.

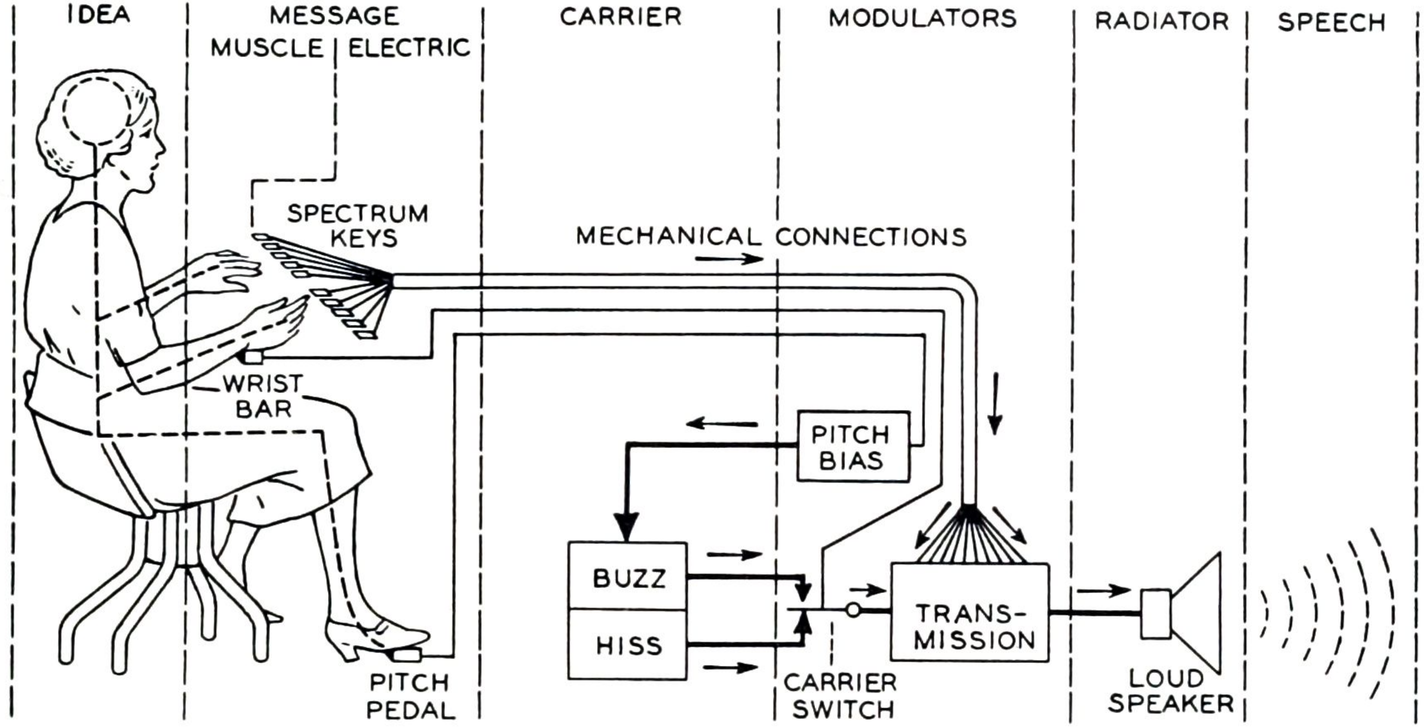
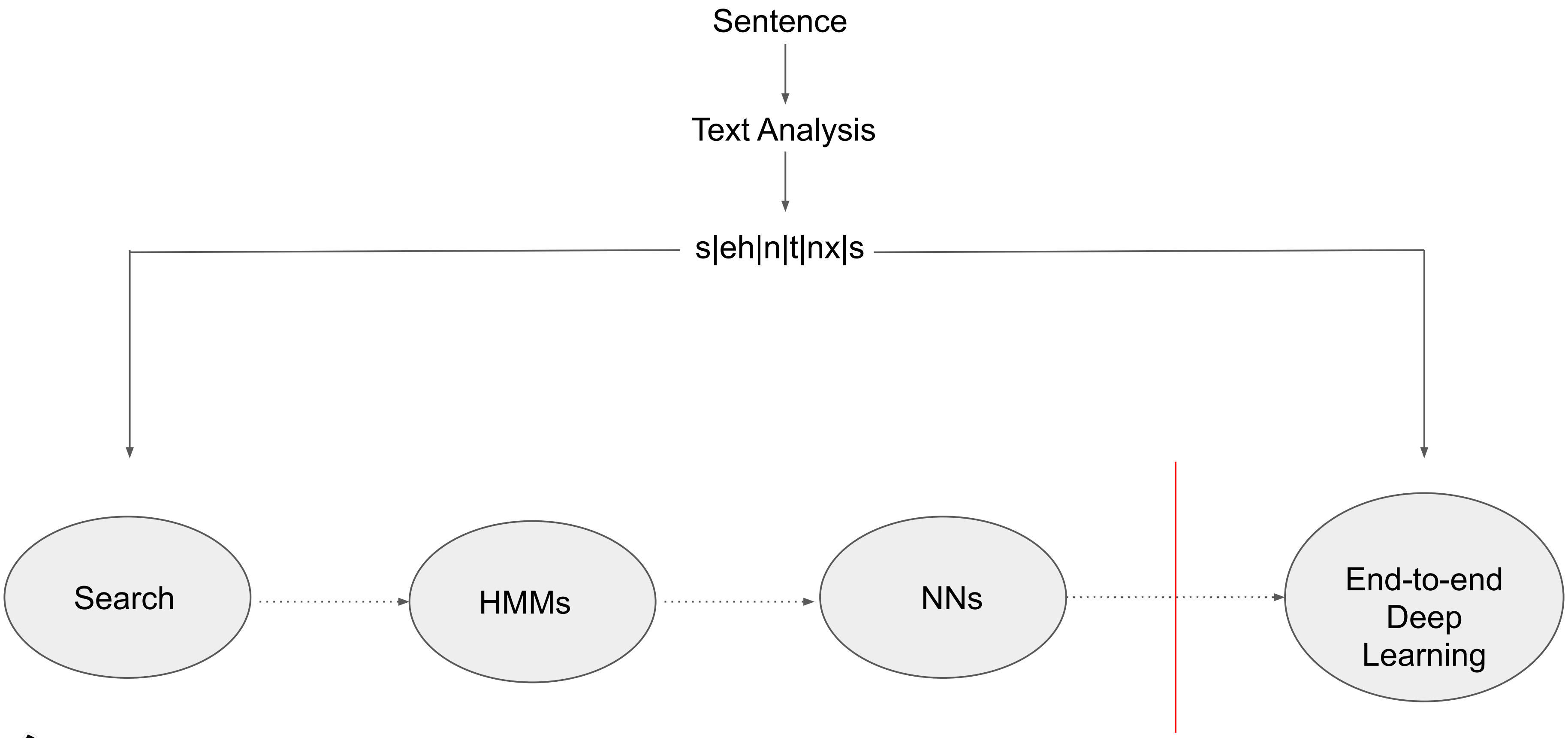


Fig. 8—Schematic circuit of the voder.

Journey

- 1939: The Voder - first electronic voice synthesiser
    - 🔊 🔊
- 1980: Formant Based TTS system
    - 🔊
- 1990-2017: Concatenative/Hybrid(+SPSS) TTS
    - 🔊
- 2018-now: End-to-End Deep Learning TTS
    - 🔊

“If you don’t have the sound, you can’t put it out”



Sentence

Text Analysis

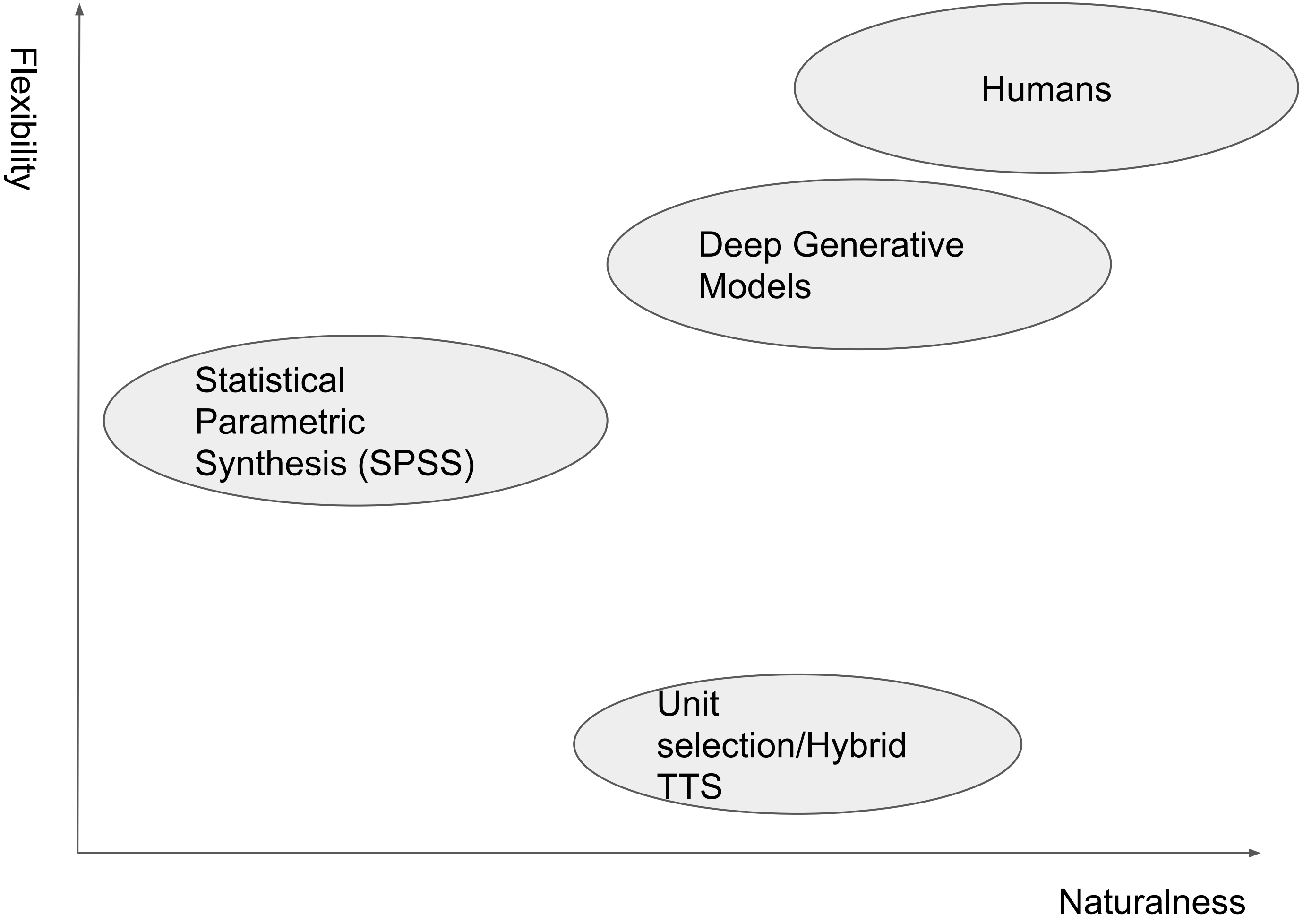
s|eh|n|t|nx|s

Search

HMMs

NNs

End-to-end Deep Learning


Flexibility
Humans
Deep Generative Models
Statistical Parametric Synthesis (SPSS)
Unit selection/Hybrid TTS
Naturalness

Thanks for tuning in!