An overview of speech synthesis: From formant to deep learning approaches

Adriana STAN

UAIC, 7.04.2021
Hello!

Adriana STAN, PhD
Assistant professor
Communications Department, Technical University of Cluj-Napoca
www.adriananastan.com
adriana.stan@com.utcluj.ro
Overview

- Primer on speech processing
- Overview of text-to-speech synthesis
- Formant synthesis
- Concatenative synthesis
- Statistical parametric speech synthesis
- Deep neural synthesis
Primer on speech processing
What is speech?

![Diagram of the vocal tract](image)

- **Nasal cavity**
- **Mouth**
- **Epiglottis**
- **Supraglottis**
- **Cartilage** (vocal cords are behind cartilage)
- **Glottis**
- **Subglottis**
- **Trachea** (windpipe)
- **Esophagus**

© Adriana STAN, 2021, no redistribution, no re-use
What is speech?

- Vowels/consonants
What is speech?

- Vowels/consonants
What is speech?

Variability

© Adriana STAN, 2021, no redistribution, no re-use
Signal processing

Analog to digital conversion
Signal processing

Stationarity
Signal processing

Quasi-periodicity
Signal processing

Spectral analysis
Signal processing

Spectral analysis
Modelling speech

- Source-filter model
- Cepstrum
- Linear Predictive Coding
- Mel frequency
Text-to-speech synthesis
Text-to-speech synthesis

- Text processing
- Acoustic modelling

© Adriana STAN, 2021, no redistribution, no re-use

UAIC, 7.04.2021
Technical issues

- the written language
- the human ear
- speaker and speech variability
- co-articulation effects
- no fixed set of measurable factors
- no objective measure of quality
Text processing

- Normalisation: "123" -> "one hundred twenty three"
- Tokenisation: "This is a question? Or an answer."
- Phonetic transcription: "hello" -> "/həˈləʊ/"
- Lexical stress assignment: "record" -> "reˈcord" or "recoˈrd"
- POS tagging: "play" -> verb
- Syllabification: "resume" -> "re.su.me"
Rule-based synthesis
Formant synthesis

Figure 2.2: Klaat’s formant synthesiser (after [Benesty et al., 2007])
Articulatory synthesis

VocalTractLab, https://www.vocaltractlab.de/
Corpus-based synthesis
Concatenative synthesis

Speech Zone, http://speech.zone/interactive-unit-selection/

© Adriana STAN, 2021, no redistribution, no re-use  
UAIC, 7.04.2021
HMM-based synthesis
HMM-based synthesis

Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W Black, Keiichi Tokuda, The HMM-based speech synthesis system (HTS) version 2.0, SSW, 2007
Deep neural synthesis
Parametric neural synthesis

Zhizheng Wu, Oliver Watts, Simon King, "Merlin: An Open Source Neural Network Speech Synthesis System" in Proc. 9th ISCA Speech Synthesis Workshop (SSW9), September 2016, Sunnyvale, CA, USA.
End-to-end speech synthesis

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu, *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*, ICASSP 2018
End-to-end speech synthesis

End-to-end speech synthesis

End-to-end speech synthesis

Neural vocoders


© Adriana STAN, 2021, no redistribution, no re-use
Neural vocoders

Where next?
Open problems

- one-shot inference and speaker adaptation
- spontaneous speech
- adequate prosody
- affective rendering
Speech synthesis challenges

- Blizzard Challenge: http://www.festvox.org/blizzard/
Quality measurements

- Naturalness
- Intelligibility
- Speaker similarity
- (Prosodic adequacy)
Thank you for your attention!

Questions?

adriana.stan@com.utcluj.ro
Bibliography

- Adriana STAN, Mircea GIURGIU, Prelucrarea semnalului vocal folosind Python, UTPress, 2021
- Adriana STAN, Beata LORINCZ, Generating the Voice of the Interactive Virtual Assistant, IntechOpen, 2021
- Paul Taylor, Text-to-speech synthesis, Cambridge University Press, 2009
- Simon King, Speech Zone, online
- PapersWithCode: https://paperswithcode.com/methods/area/audio