Source: https://lingojam.com/StephenHawkingVoiceGenerator

Chair of Real-Time Computer Systems
TUM Department of Electrical and Computer Engineering
Technical University of Munich

Final Presentation of Advanced Seminar

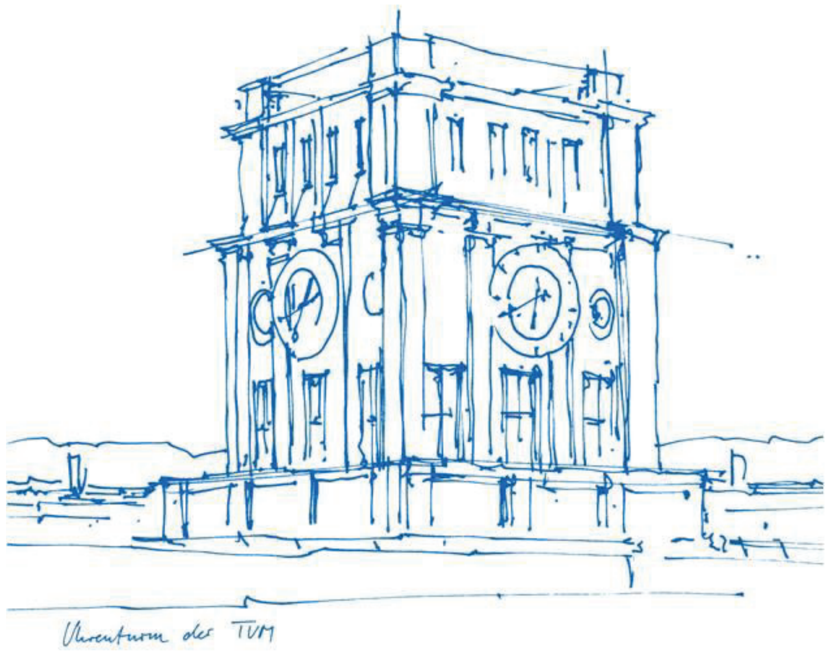
Hannes Bohnengel

Technical University of Munich

TUM Department of Electrical and Computer Engineering

Chair of Real-Time Computer Systems

Munich, 21 July 2017



The Impact of Deep Learning on Speech Synthesis with Mobile Devices

Outline

1. Content of Paper

2. Speech Synthesis in General

3. Introducing Deep Learning Models

4. Speech Synthesis on Mobile Devices

5. Conclusions

Content of Paper

1. **Introduction**
2. **Conventional Speech Synthesis**
   a) Motivation and Approaches
   b) HMM-based Synthesis → Black et al. (2007), ICASSP '07
3. **Speech Synthesis with Deep Learning Models**
   a) One Specific Approach for Improvement → Zen et al. (2013), ICASSP '13
   b) Other Ways for Improvement → Hashimoto et al. (2015), ICASSP '13
4. **Speech Synthesis on Mobile Devices**
   a) Motivation and Challenges
   b) Optimized HMM-based Synthesis → Tóth et al. (2012), JACIII '12
   c) Deep Learning-based Synthesis → Boroş et al. (2015), MEDES '15
5. **Conclusions**

Outline

1. Content of Paper
2. Speech Synthesis in General
3. Introducing Deep Learning Models
4. Speech Synthesis on Mobile Devices
5. Conclusions

Typical Applications of Speech Synthesis

Navigation Systems

Telephone-based Dialogue Systems

Reading Aid

Speech-to-Speech Translation

Voice Communication Aid
(Stephen Hawking)

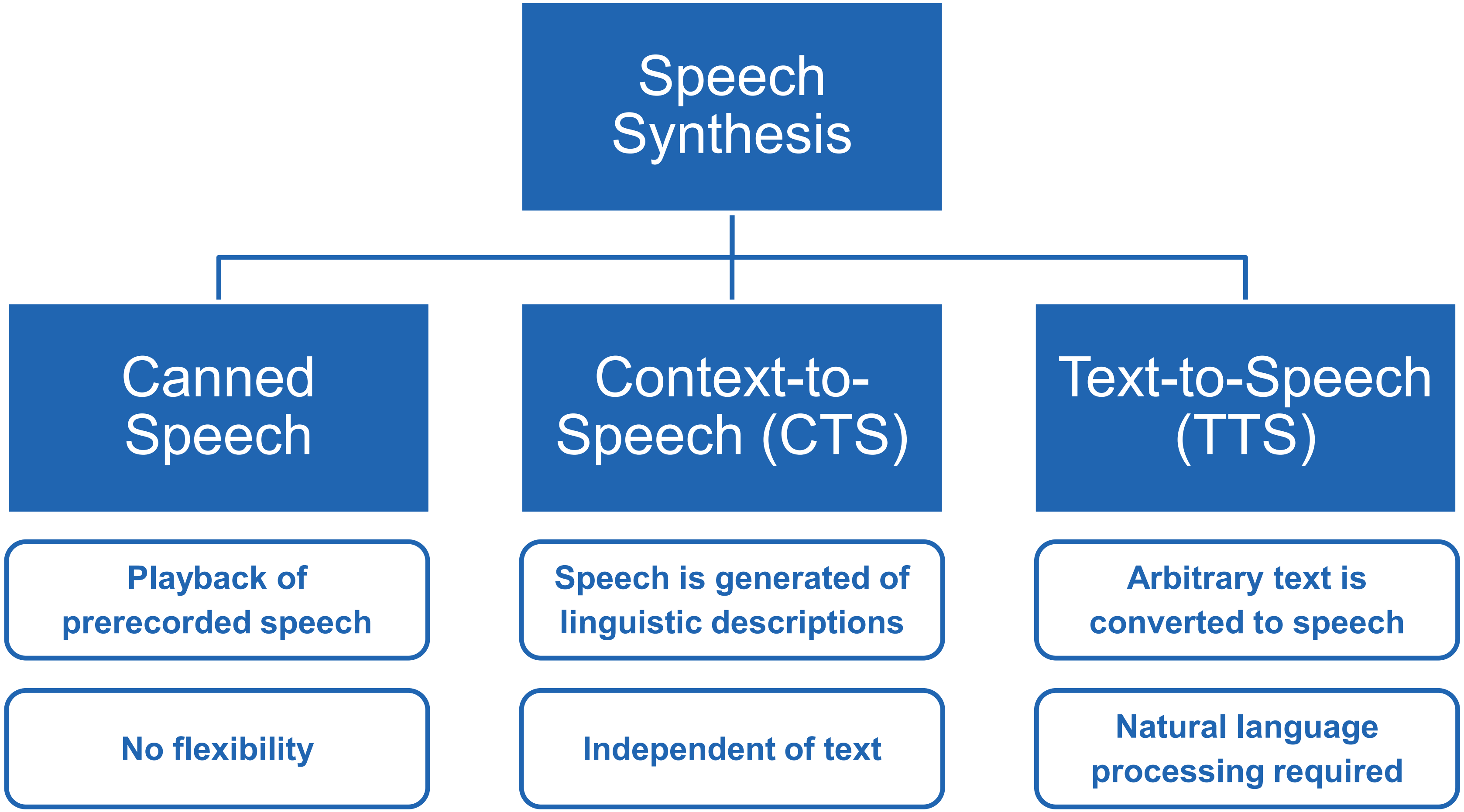
Virtual Assistants

Voice cloning

Public Announcements

Communication in Air Traffic

Types of Speech Synthesis



Source: Own visualization

Text-to-Speech – Overview

Front-end

Natural Language Processing

- Part-of-speech tagging
- Text normalization
- Phonetic transcription
- Syllabification
- Stress prediction
- Prosodic analysis

Back-end

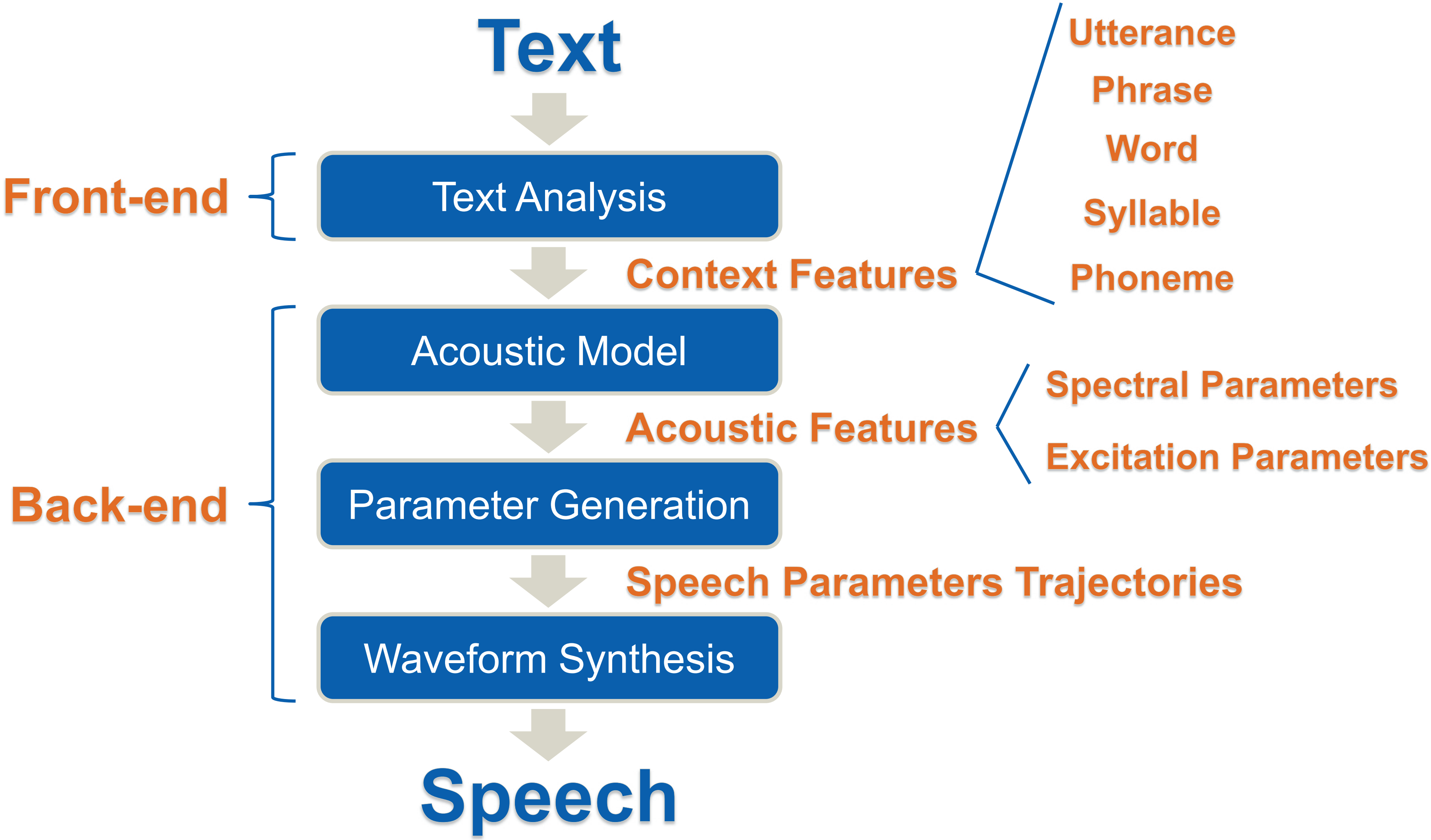
Digital Signal Processing

Depends on synthesis model

- Parametric
- Concatenative
- Statistical parametric

Source: Boroș et al. (2015) Robust deep-learning models for text-to-speech synthesis support on embedded devices, MEDES'15

Text-to-Speech – Function blocks



Source: Own visualization

Text-to-Speech – Synthesis Models

Table 1: Comparison of different speech synthesis techniques

Technique	Advantages	Drawbacks
Formant-based (Parametric)	Very small footprint	Very artificial and metallic voice
Unit-selection (Concatenative)	Very high voice quality possible	Large database required
HMM-based (Statistical parametric)	Adjustable voice and small footprint	Voice sounds muffled

Source: Own visualization

Sample of HMM-based speech

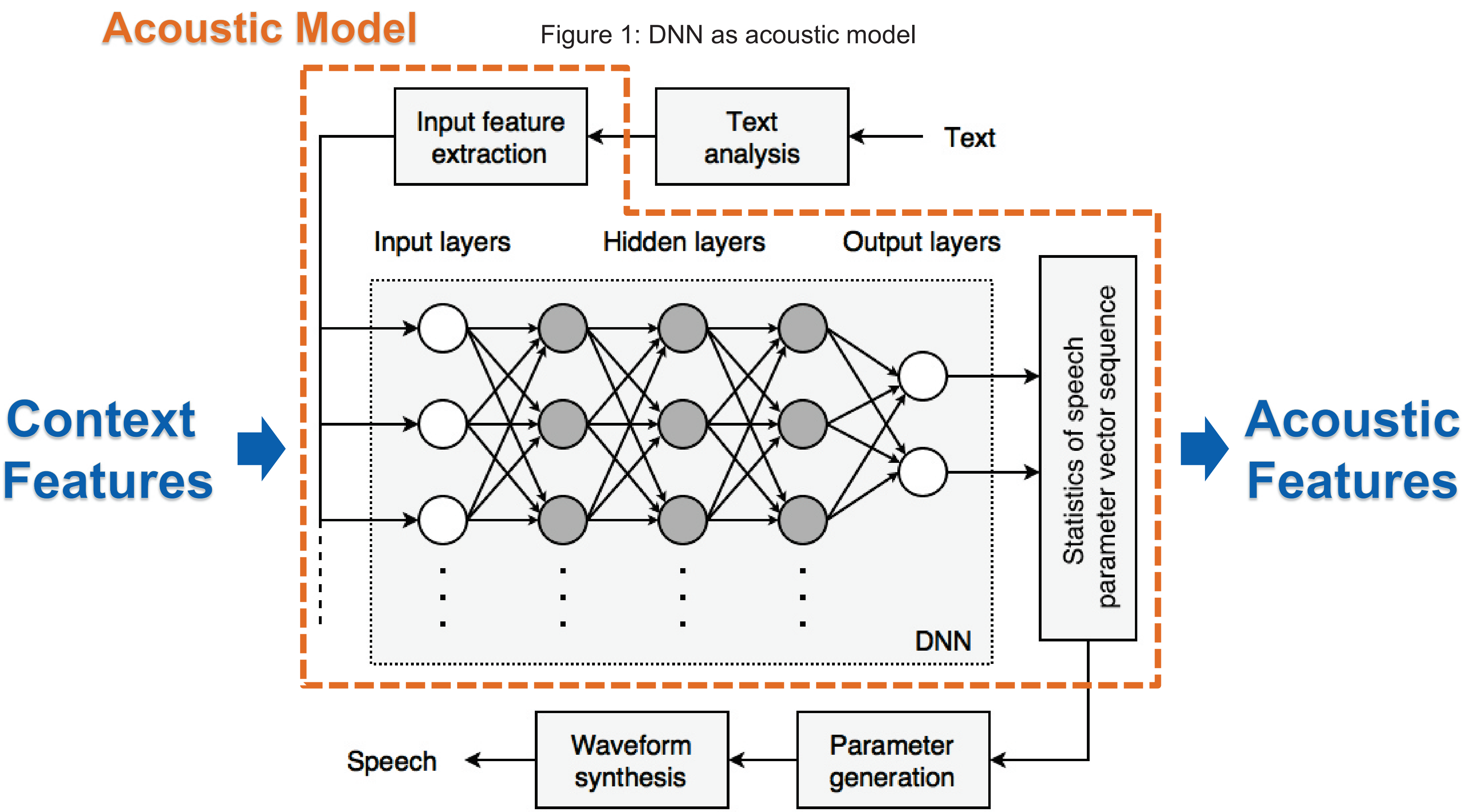
Source: http://flite-hts-engine.sp.nitech.ac.jp/index.php

Outline

1. Content of Paper
2. Speech Synthesis in General
3. Introducing Deep Learning Models
4. Speech Synthesis on Mobile Devices
5. Conclusions

Introducing Deep Learning Models

Zen et al. (2013), ICASSP ’13



Source: Zen et al. (2013) Statistical parametric speech synthesis using deep neural networks, ICASSP’13

Results of Experiments

Objective evaluation

→ DNN-based systems have less distortion

→ HMM-based systems have a lower error rate in some cases

Subjective evaluation

→ DNN-based systems are preferred

→ Described as less muffled

Table 2: Subjective scores

HMM-based (scaling factor)	DNN-based (neurons per layer)	Neutral
15.8 % (16)	38.5 % (256)	45.7 %
16.1 % (4)	27.2 % (512)	56.8 %
12.7 % (1)	36.6 % (1024)	50.7 %

Source: Zen et al. (2013) Statistical parametric speech synthesis using deep neural networks, IEEE International Conference on Acoustics, Speech and Signal Processing

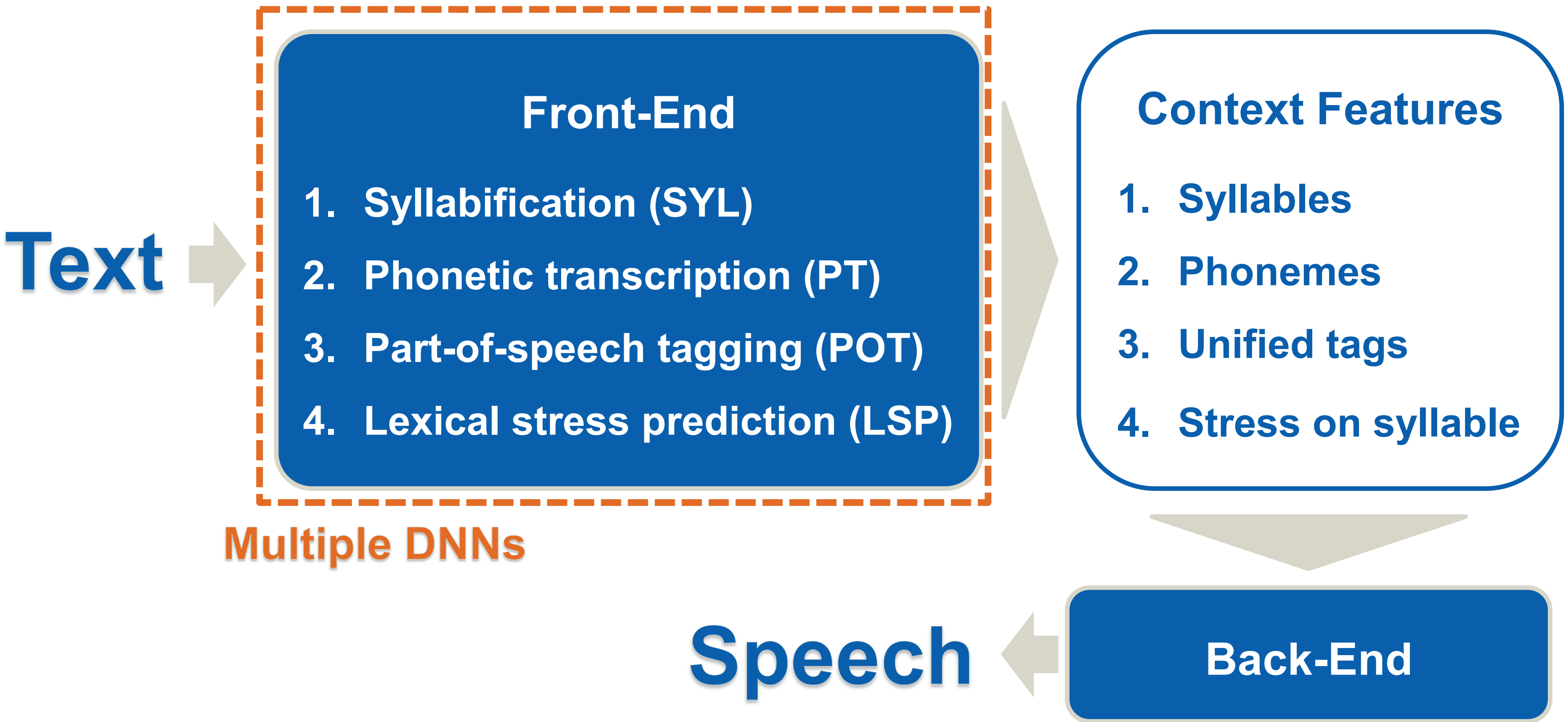
Outline

1. Content of Paper
2. Speech Synthesis in General
3. Introducing Deep Learning Models
4. Speech Synthesis on Mobile Devices
5. Conclusions

Speech Synthesis on Mobile Devices

Boroş et al. (2015), MEDES ’15

→ Introducing a DNN into the front-end of a TTS-System to decrease the model size



Source: Own visualization

Results of Experiments

Table 3: Resulting accuracy and footprint size

	SYL		PT		POT		LSP	
	Con.	DNN	Con.	DNN	Con.	DNN	Con.	DNN
Accuracy	99.01 %	98.23 %	96.29 %	96.16 %	98.19 %	95.16 %	98.80 %	97.67 %
Size	9.4 MB	36.7 KB	1.4 MB	43.4 KB	96 MB	178 KB	6 MB	110 KB

Source: Boroș et al. (2015) Robust deep-learning models for text-to-speech synthesis support on embedded devices, MEDES ’15

Overall reduction of model size by ~ 60 %

Con. = Conventional

Outline

1. Content of Paper
2. Speech Synthesis in General
3. Introducing Deep Learning Models
4. Speech Synthesis on Mobile Devices
5. Conclusions

Conclusions

Speech synthesis is an important technology

→ Huge research volume

→ Practical relevance with many application

Deep learning models have emerged in the last decade

→ Strength: Mapping complex input features to simple output features

→ Deep learning can be used to improve speech synthesis

Huge number of mobile devices

→ Need for robust and resource-efficient implementations

→ Deep learning models can be used to achieve this