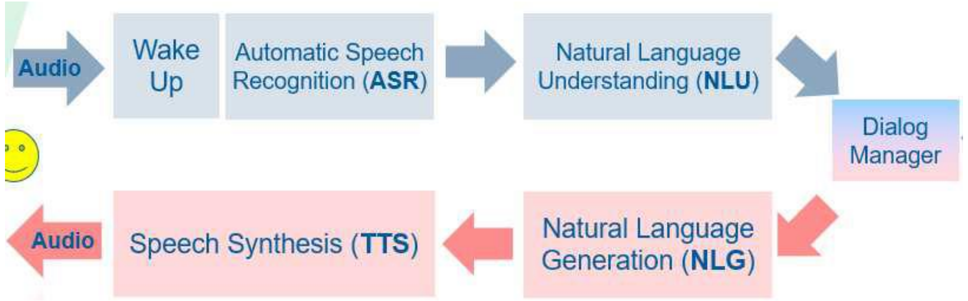
Text-to-Speech

Motivation

Text-to-Speech

Motivation

- **Used in voice Assistants**


Audio
Wake Up
Automatic Speech Recognition (ASR)
Natural Language Understanding (NLU)
Dialog Manager
Natural Language Generation (NLG)
Speech Synthesis (TTS)
Audio

Stephen Hawking speaks at MIT (YouTube): *https://youtu.be/b-2GV0T5Zpc?t=130*
Klatt‘s Last Tapes - History of Speech Synthesis - Radio 4: https://youtu.be/097K1uMIPyQ?t=1143
Read webpages or PDFs: e.g. https://ttsreader.com/

Text-to-Speech

Motivation

- **Used in voice Assistants**
- **Turn ebooks into audiobooks**

Stephen Hawking speaks at MIT (YouTube): *https://youtu.be/b-2GV0T5Zpc?t=130*
Klatt‘s Last Tapes - History of Speech Synthesis - Radio 4: https://youtu.be/097K1uMIPyQ?t=1143
Read webpages or PDFs: e.g. https://ttsreader.com/

Text-to-Speech

Motivation

- **Used in voice Assistants**
- **Turn ebooks into audiobooks**
- **Assistive communication technology for people with**

Stephen Hawking speaks at MIT (YouTube): *https://youtu.be/b-2GV0T5Zpc?t=130*
Klatt‘s Last Tapes - History of Speech Synthesis - Radio 4: https://youtu.be/097K1uMIPyQ?t=1143
Read webpages or PDFs: e.g. https://ttsreader.com/

Text-to-Speech

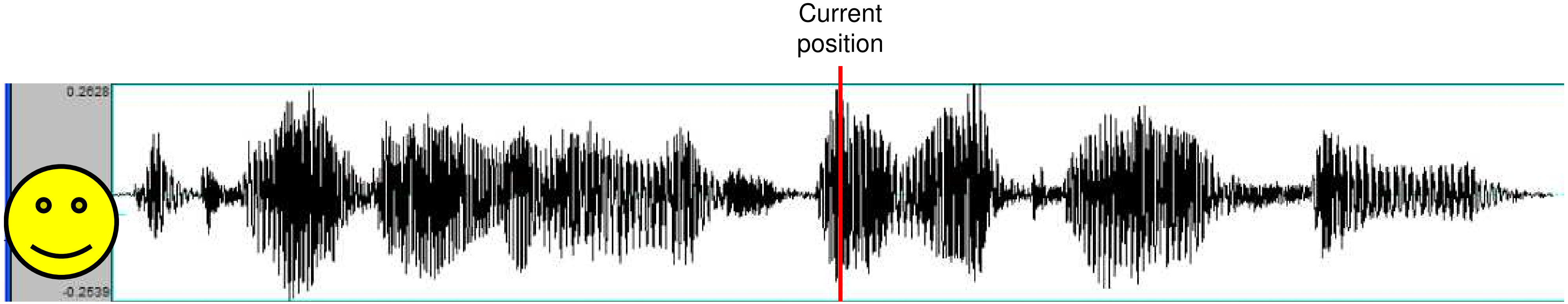
Motivation

- **Used in voice Assistants**
- **Turn ebooks into audiobooks**
- **Assistive communication technology for people with**
  - **Visual impairments**
  - **Reading difficulties**
  - **speaking disorders**

Stephen Hawking speaks at MIT (YouTube): *https://youtu.be/b-2GV0T5Zpc?t=130*
Klatt‘s Last Tapes - History of Speech Synthesis - Radio 4: https://youtu.be/097K1uMIPyQ?t=1143
Read webpages or PDFs: e.g. https://ttsreader.com/

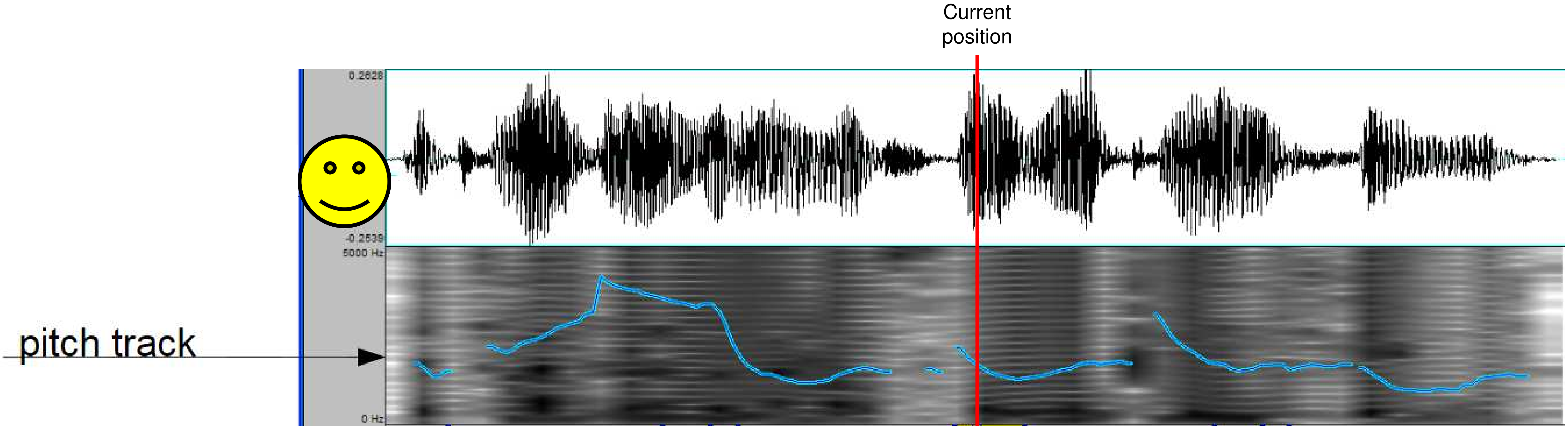
Text-to-Speech - Vocabulary

Wave forms



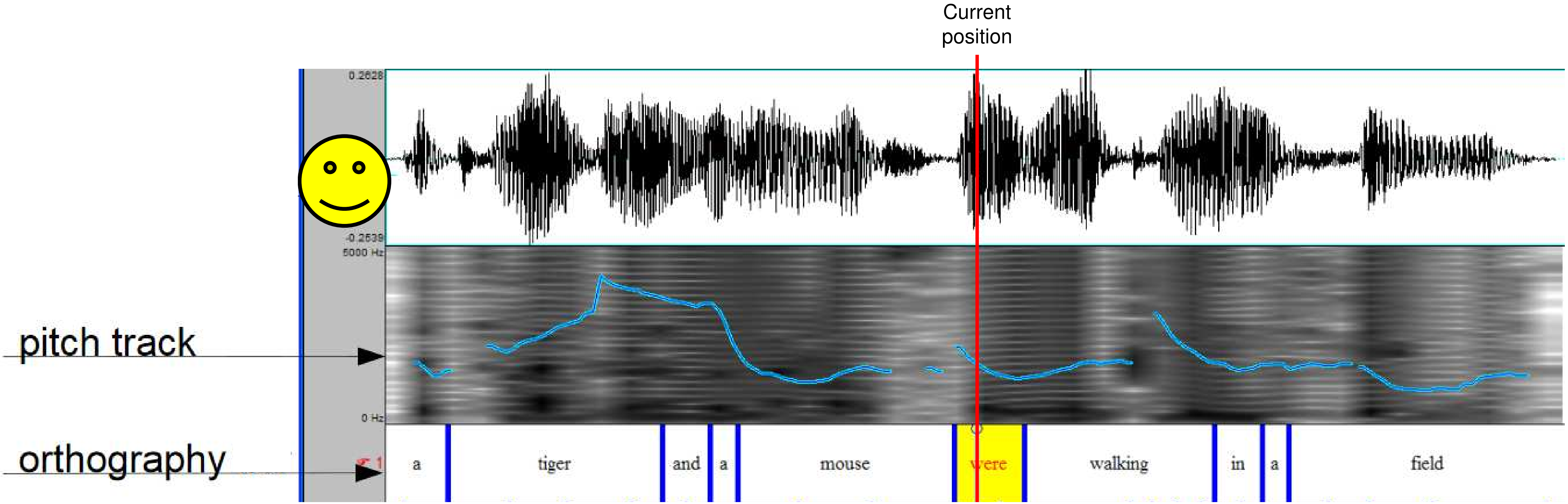
Text-to-Speech - Vocabulary

Pitch



Text-to-Speech - Vocabulary

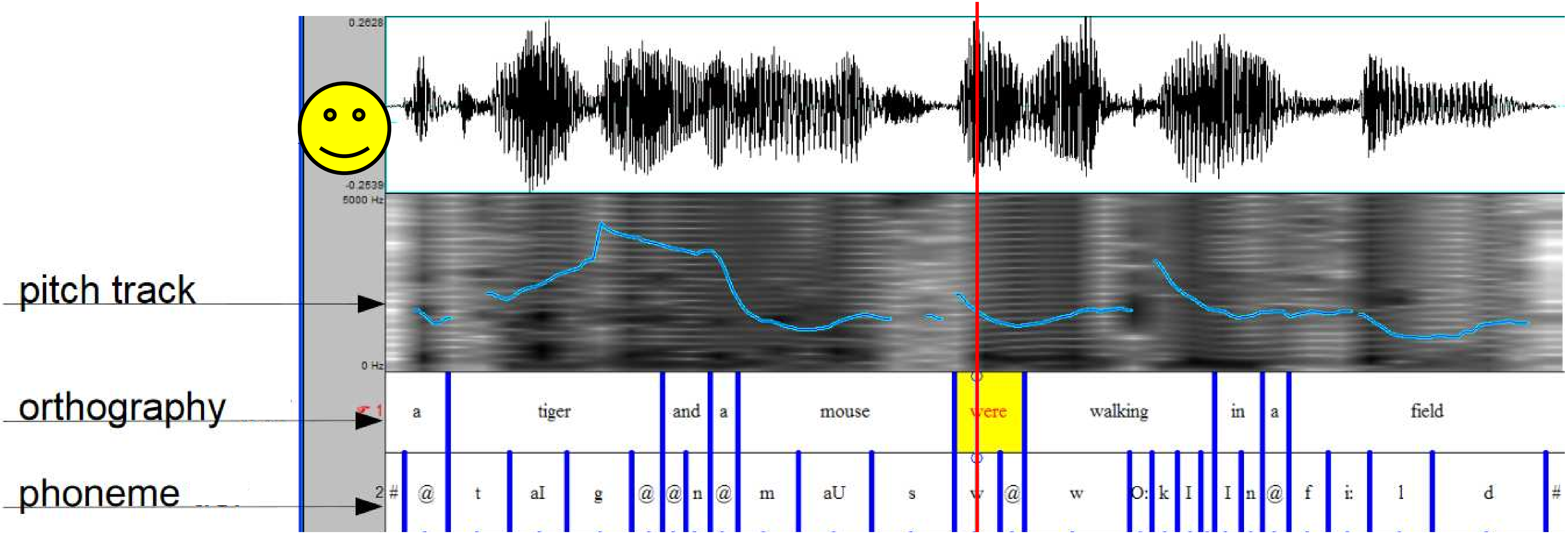
Orthography



Text-to-Speech - Vocabulary

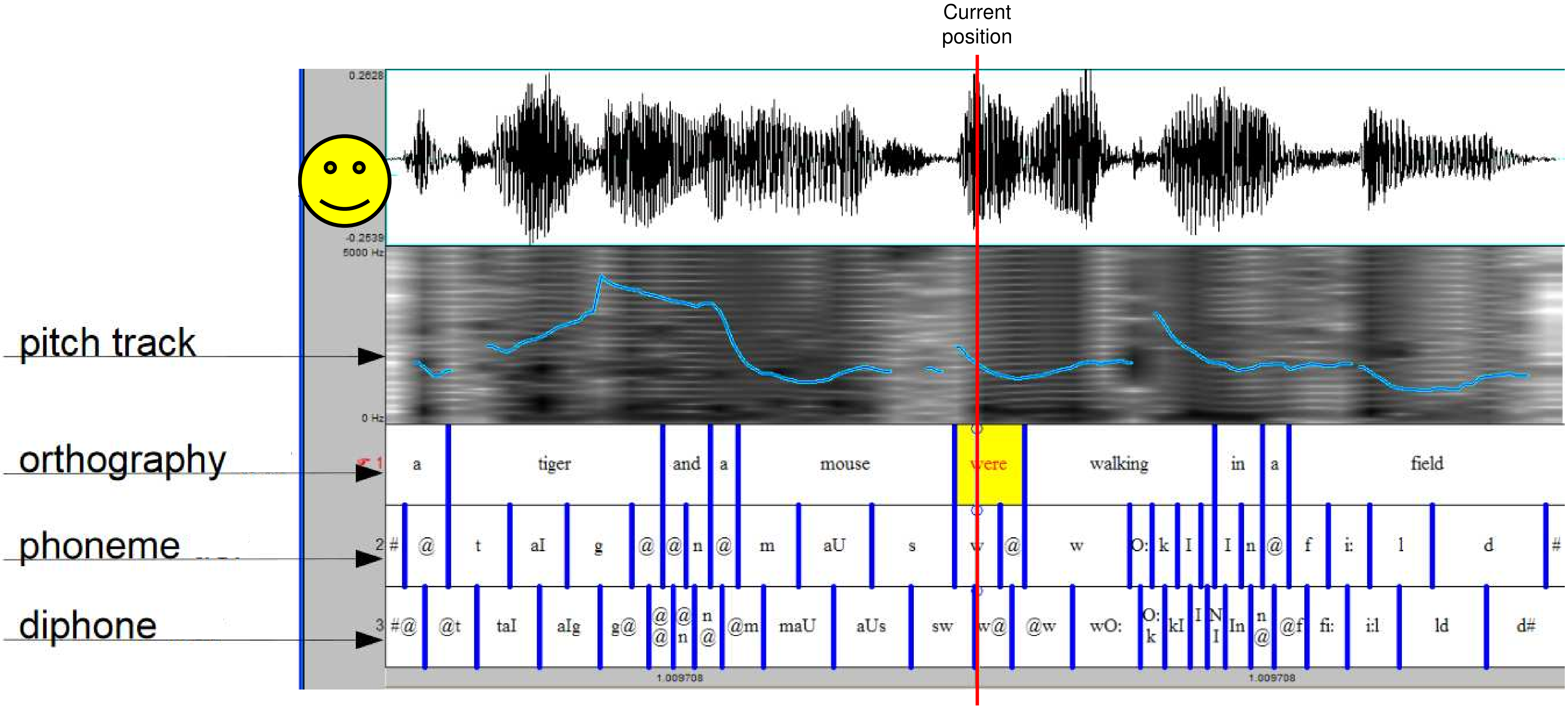
Phonemes

Current position



Text-to-Speech - Vocabulary

Diphones



Current position

pitch track

orthography

phoneme

diphone

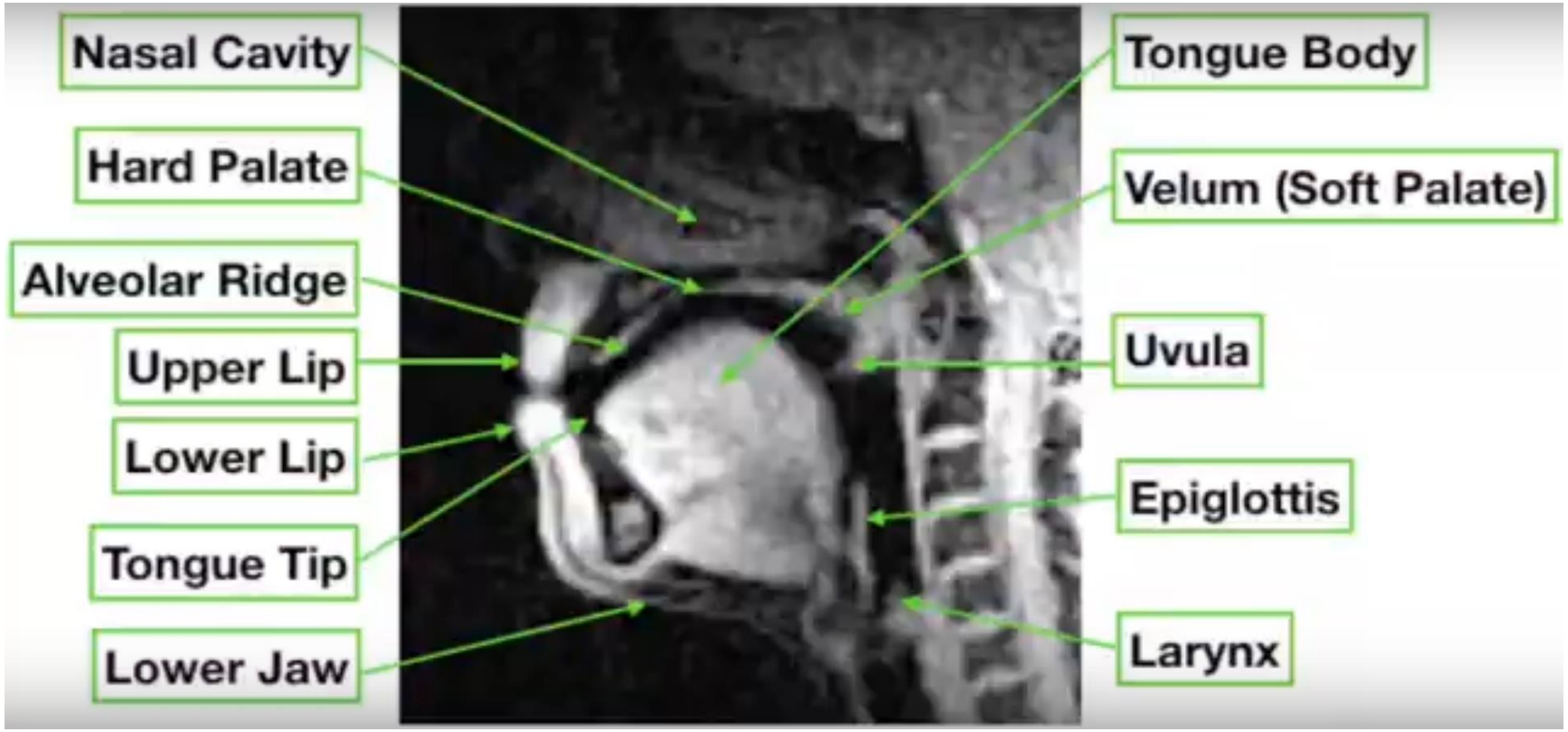
1 a tiger and a mouse were walking in a field

2 # @ t aI g @ @ n @ m aU s w @ w O: k I I n @ f i: l d #

3 #@ @t taI aIg g@ @@ @n n@ @m maU aUs sw w@ @w wO: O:k kI IN NI In n@ @f fi: i:l ld d#

Vocal Tract

MRI

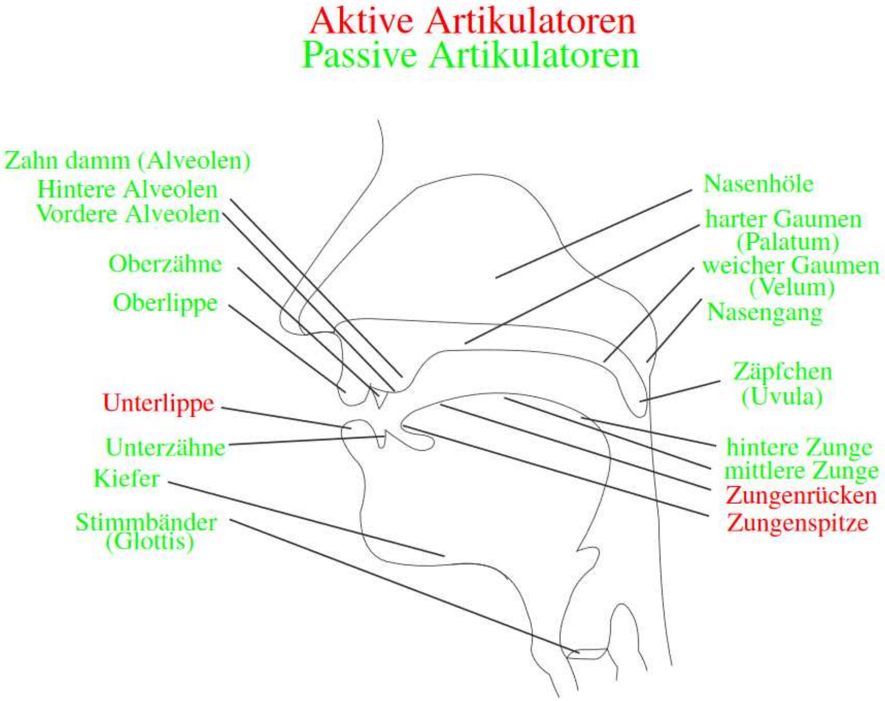


https://youtu.be/Wrbe5fH888k

Subfield of articulatory phonetics

Vocal tract & classification of sounds

The vocal tract can be well described as an all-pole filter, which can be useful, for example, for the analysis or synthesis of speech signals. The speech organs that play a special role in sound production or shaping are called articulators. A distinction is made between the more or less consciously influenced articulators and those that are only used, or between active and passive articulators. In order to describe the many, different sounds of the human language, one needs first a smallest unit, which can serve as basis for a description alphabet. In phonetics, this smallest unit is called a sound or a phon.

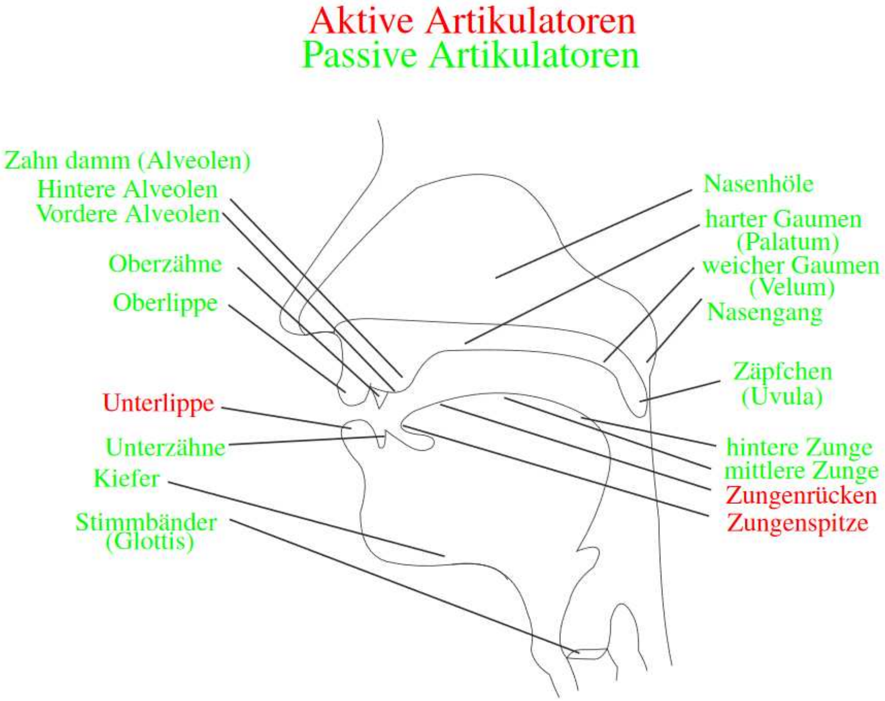


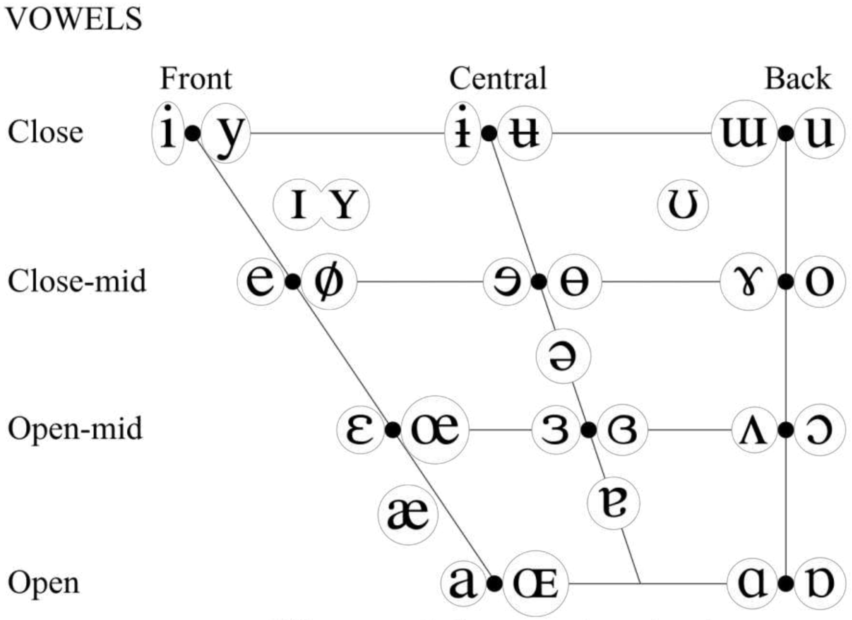
https://www.ei.ruhr-uni-bochum.de/media/ei/lehrmaterialien/spracherkennung/d0909549b9003dc1defe7f7a960ce6624d1de7f9/SkriptASE2017b.pdf

Subfield of articulatory phonetics

Vowel tract & classification of sounds

The vocal tract can be well described as an all-pole filter, which can be useful, for example, for the analysis or synthesis of speech signals. The speech organs that play a special role in sound production or shaping are called articulators. A distinction is made between the more or less consciously influenced articulators and those that are only used, or between active and passive articulators. In order to describe the many, different sounds of the human language, one needs first a smallest unit, which can serve as basis for a description alphabet. In phonetics, this smallest unit is called a sound or a phon.





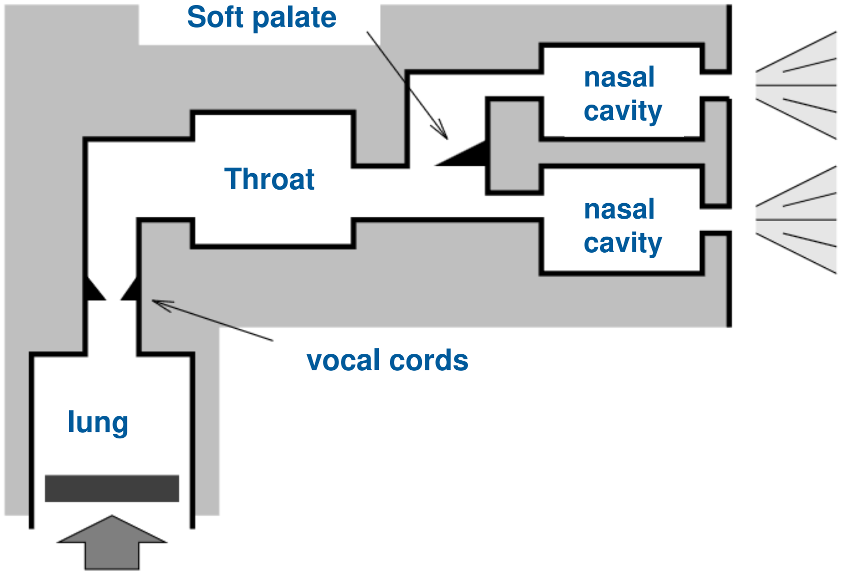
Where symbols appear in pairs, the one to the right represents a rounded vowel.

https://www.ei.ruhr-uni-bochum.de/media/ei/lehrmaterialien/spracherkennung/d0909549b9003dc1defe7f7a960ce6624d1de7f9/SkriptASE2017b.pdf

Subfield of articulatory phonetics

Physiologically motivated model of speech generation

To describe speech generation mathematically, the model in the lower left image is often used. Here, the lung serves as the source that provides the airflow for all further processes. The vocal cords determine whether the sound is to be voiced or unvoiced. In the case of unvoiced sounds, the vocal cords are so far apart that they are not influenced too much by the passing air stream; in the case of voiced sounds, they lie against each other and are moved apart at regular intervals by the air stream, thus causing them to vibrate. The frequency of this oscillation is also referred to as the fundamental frequency.

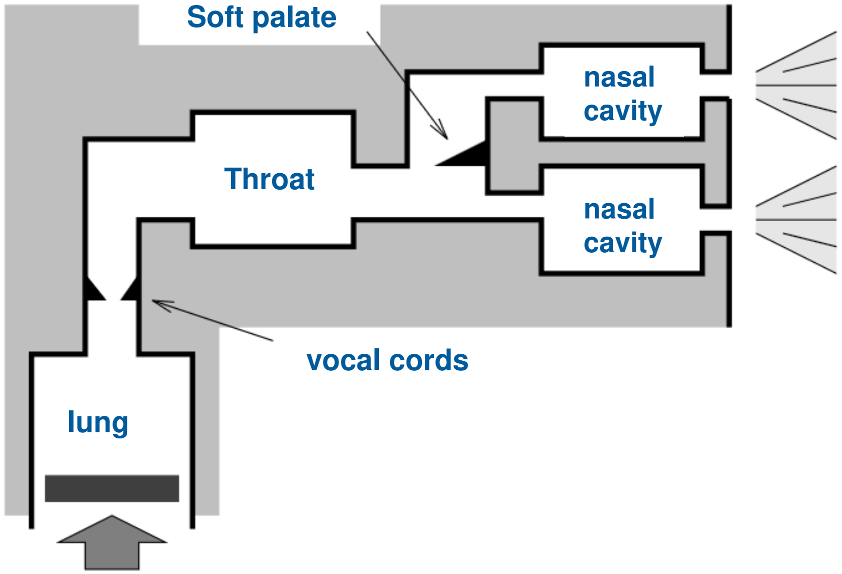


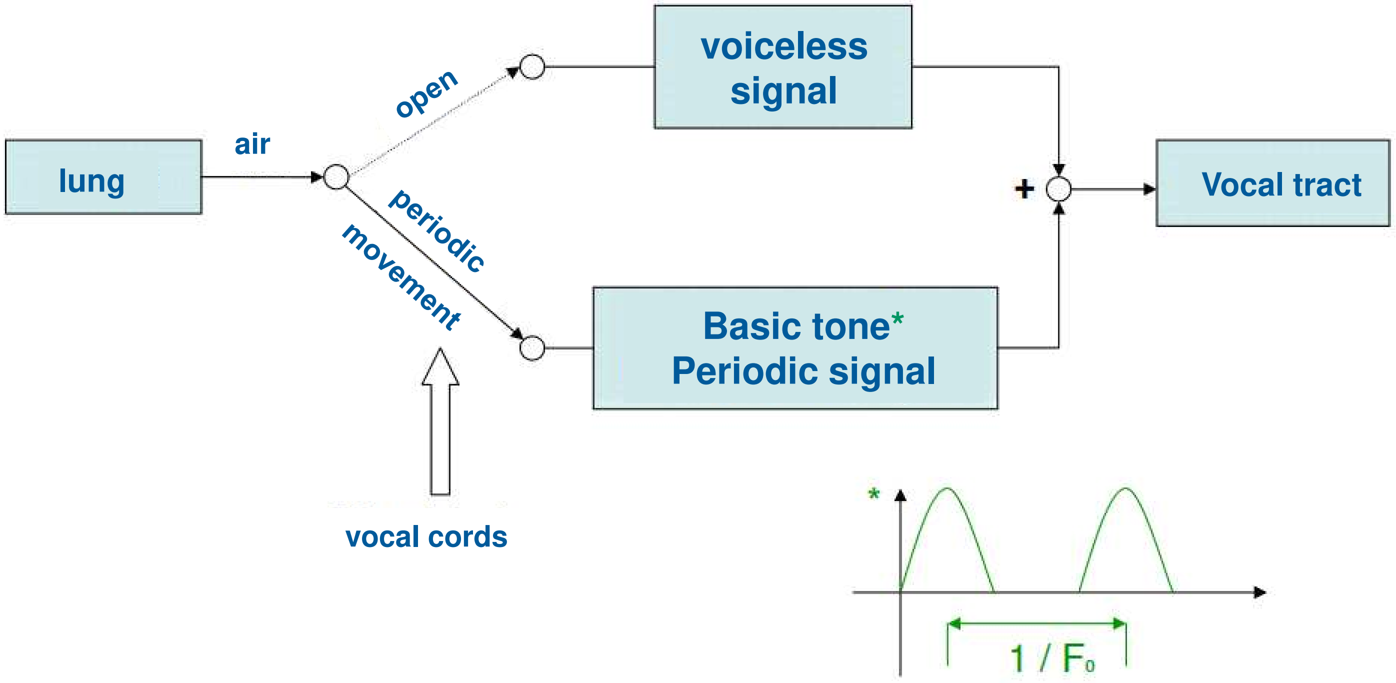
https://www.ei.ruhr-uni-bochum.de/media/ei/lehrmaterialien/spracherkennung/d0909549b9003dc1defe7f7a960ce6624d1de7f9/SkriptASE2017b.pdf

Subfield of articulatory phonetics

Physiologically motivated model of speech generation

To describe speech generation mathematically, the model in the lower left image is often used. Here, the lung serves as the source that provides the airflow for all further processes. The vocal cords determine whether the sound is to be voiced or unvoiced. In the case of unvoiced sounds, the vocal cords are so far apart that they are not influenced too much by the passing air stream; in the case of voiced sounds, they lie against each other and are moved apart at regular intervals by the air stream, thus causing them to vibrate. The frequency of this oscillation is also referred to as the fundamental frequency.

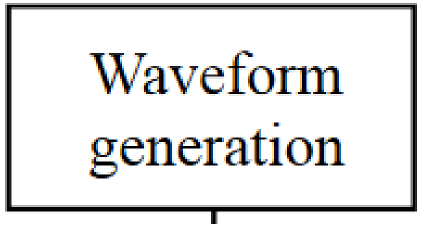




https://www.ei.ruhr-uni-bochum.de/media/ei/lehrmaterialien/spracherkennung/d0909549b9003dc1defe7f7a960ce6624d1de7f9/SkriptASE2017b.pdf

Text-to-Speech Synthesis

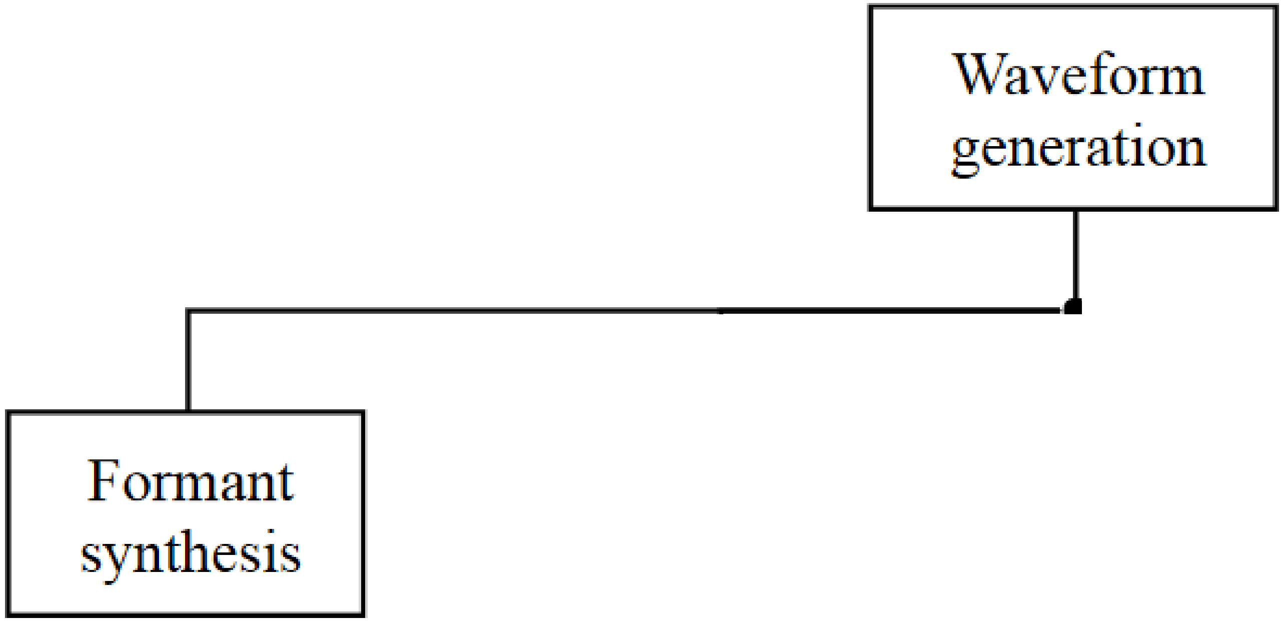
Historical Overview



Kempelen‘s speaking machine: https://www.youtube.com/watch?v=k_YUB_S6Gpo
The voder (Homer Dudley): https://www.youtube.com/watch?v=5hyI_dM5cGo

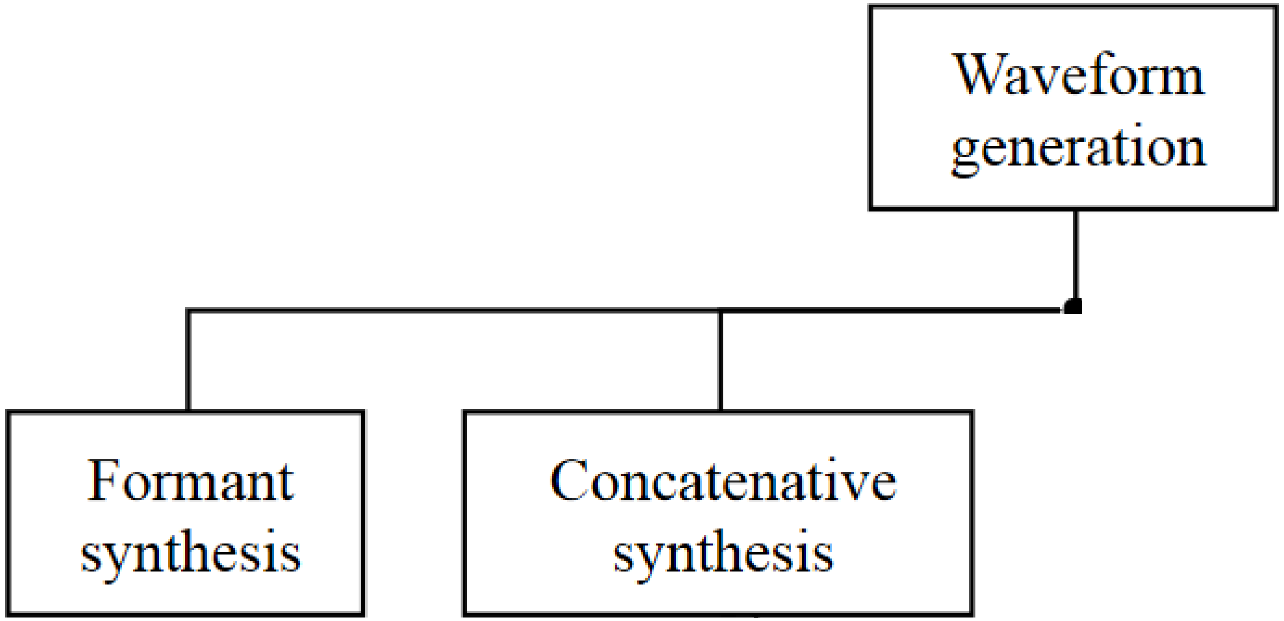
Text-to-Speech Synthesis

Historical Overview



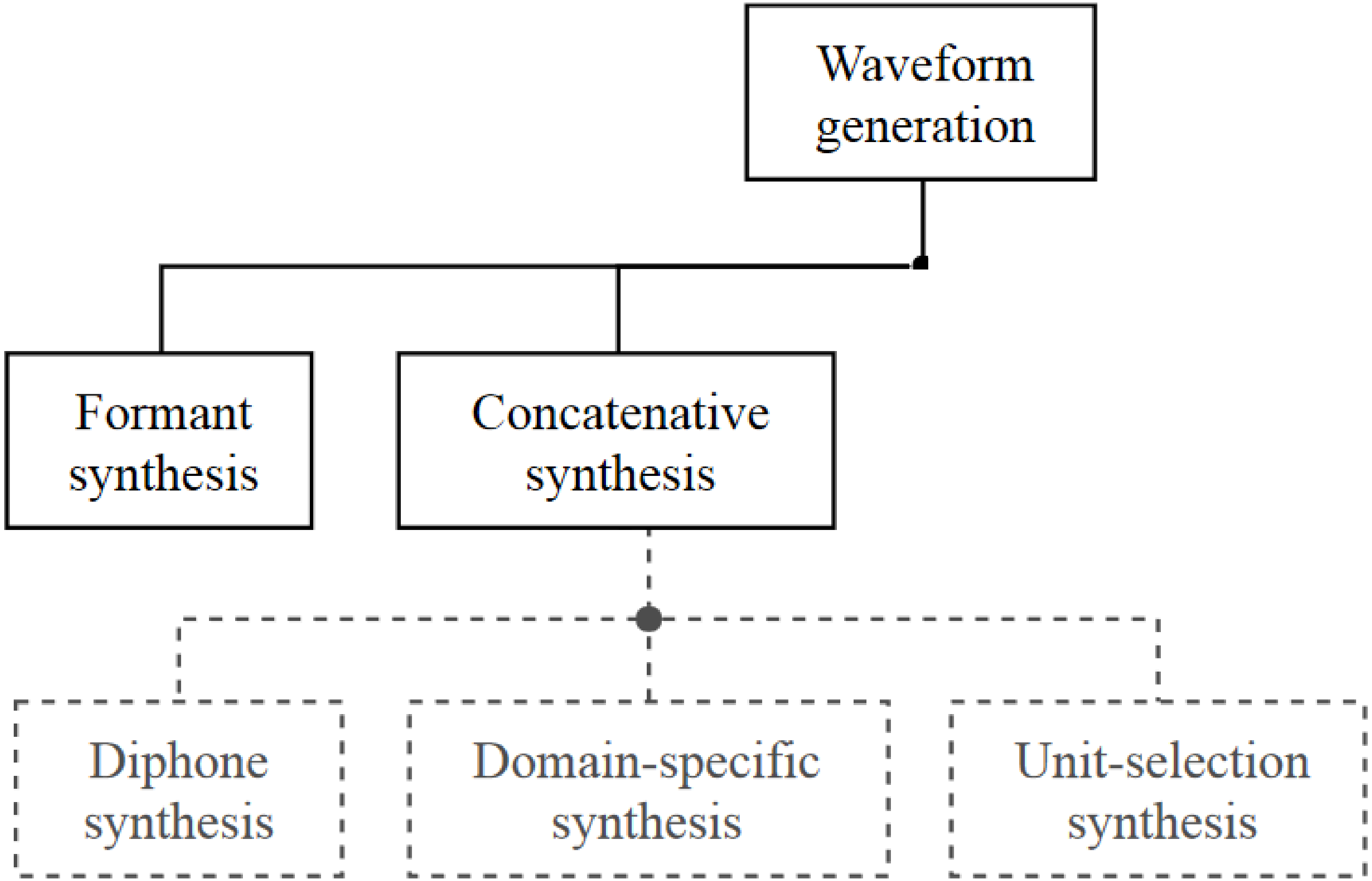
Text-to-Speech Synthesis

Historical Overview



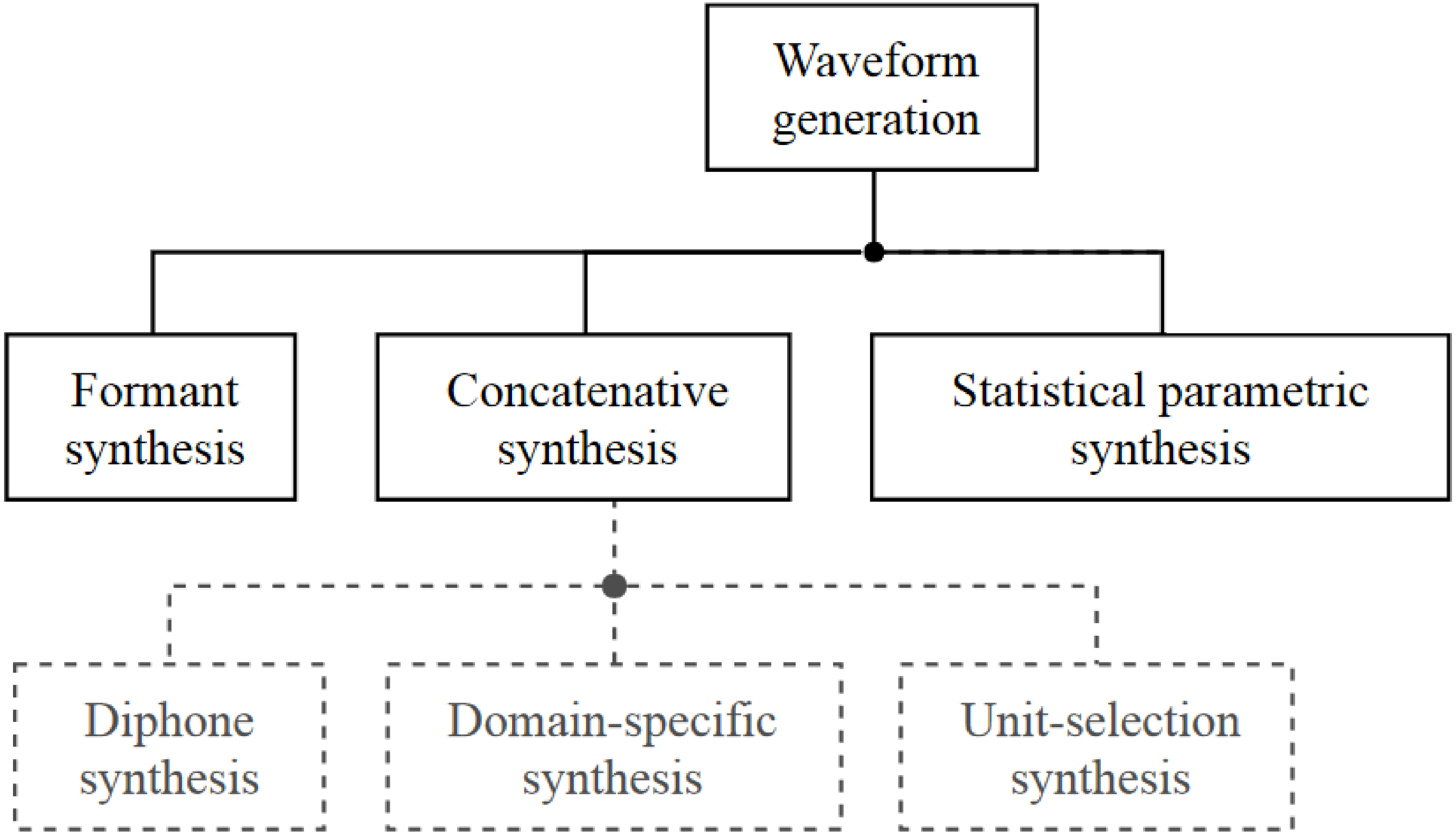
Text-to-Speech Synthesis

Historical Overview



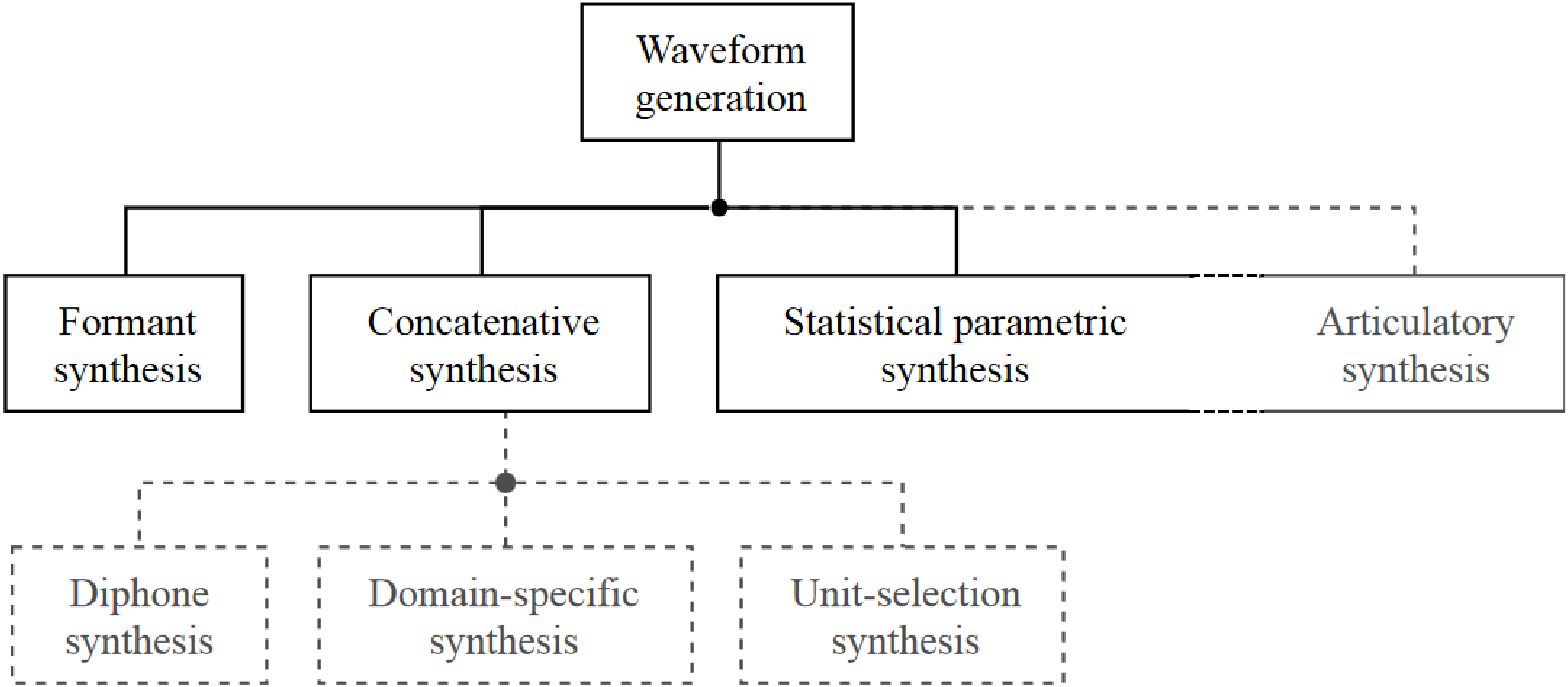
Text-to-Speech Synthesis

Historical Overview



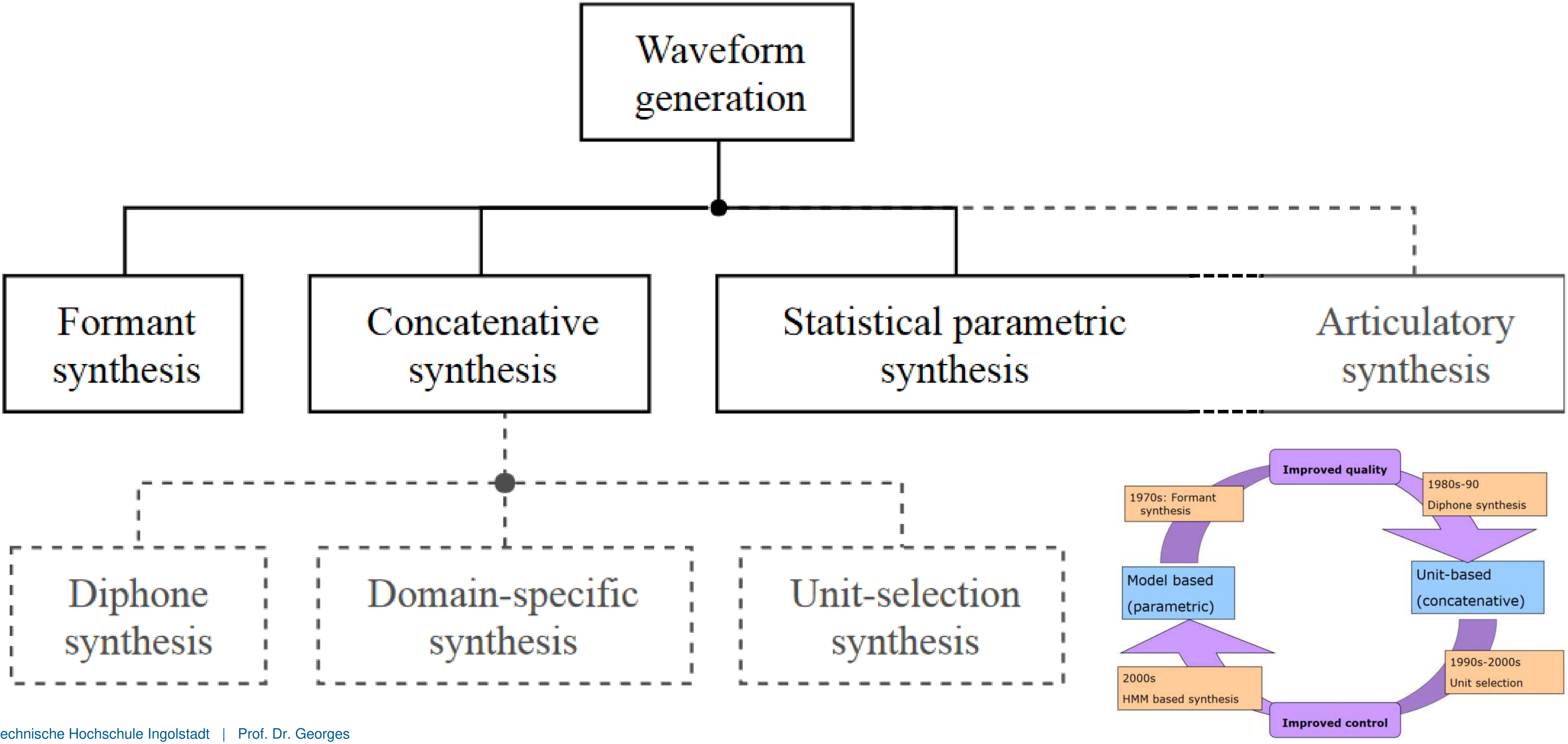
Text-to-Speech Synthesis

Historical Overview



Text-to-Speech Synthesis

Historical Overview



Waveform generation

Formant synthesis

Concatenative synthesis

Statistical parametric synthesis

Articulatory synthesis

Diphone synthesis

Domain-specific synthesis

Unit-selection synthesis

Improved quality

1980s-90 Diphone synthesis

Unit-based (concatenative)

1990s-2000s Unit selection

Improved control

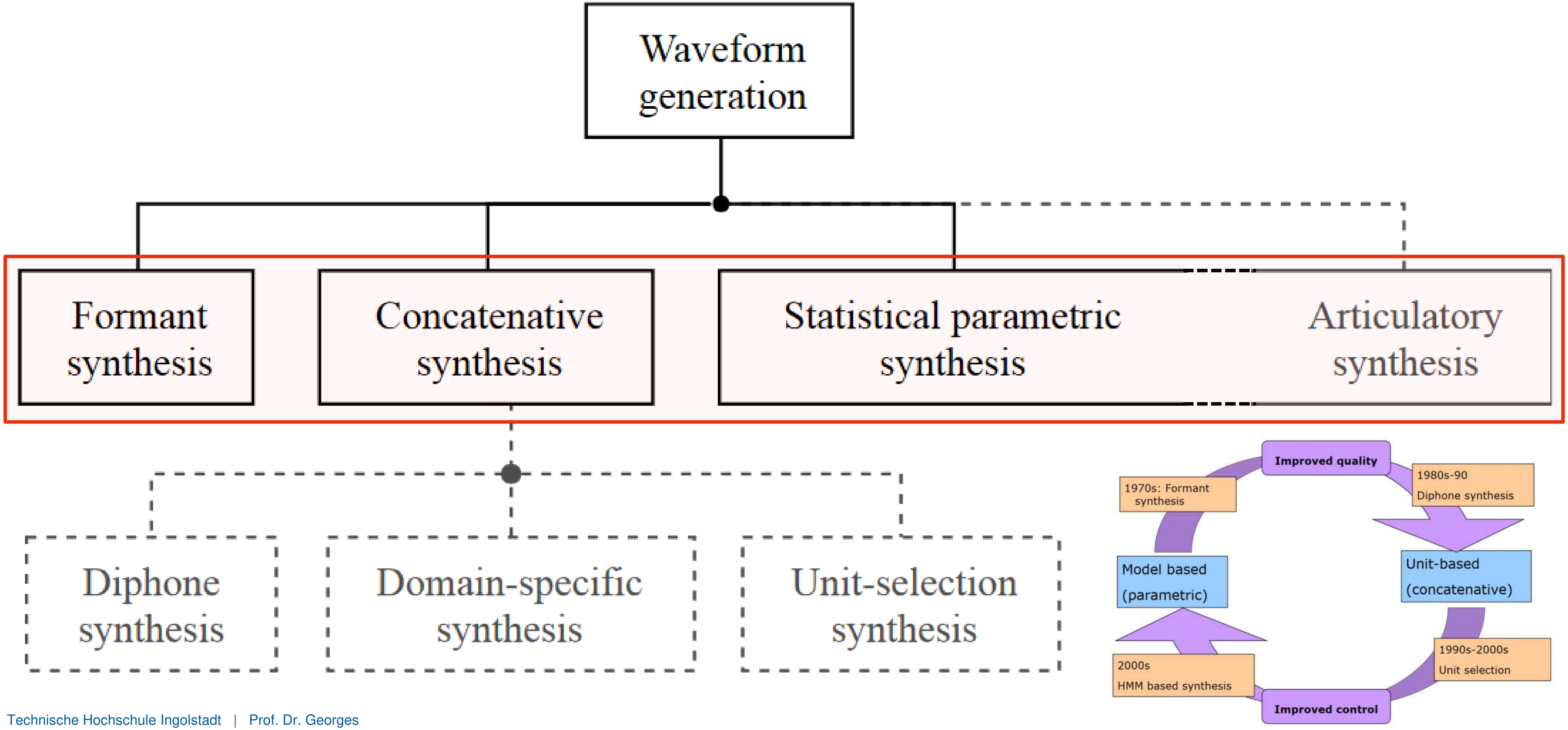
2000s HMM based synthesis

Model based (parametric)

1970s: Formant synthesis

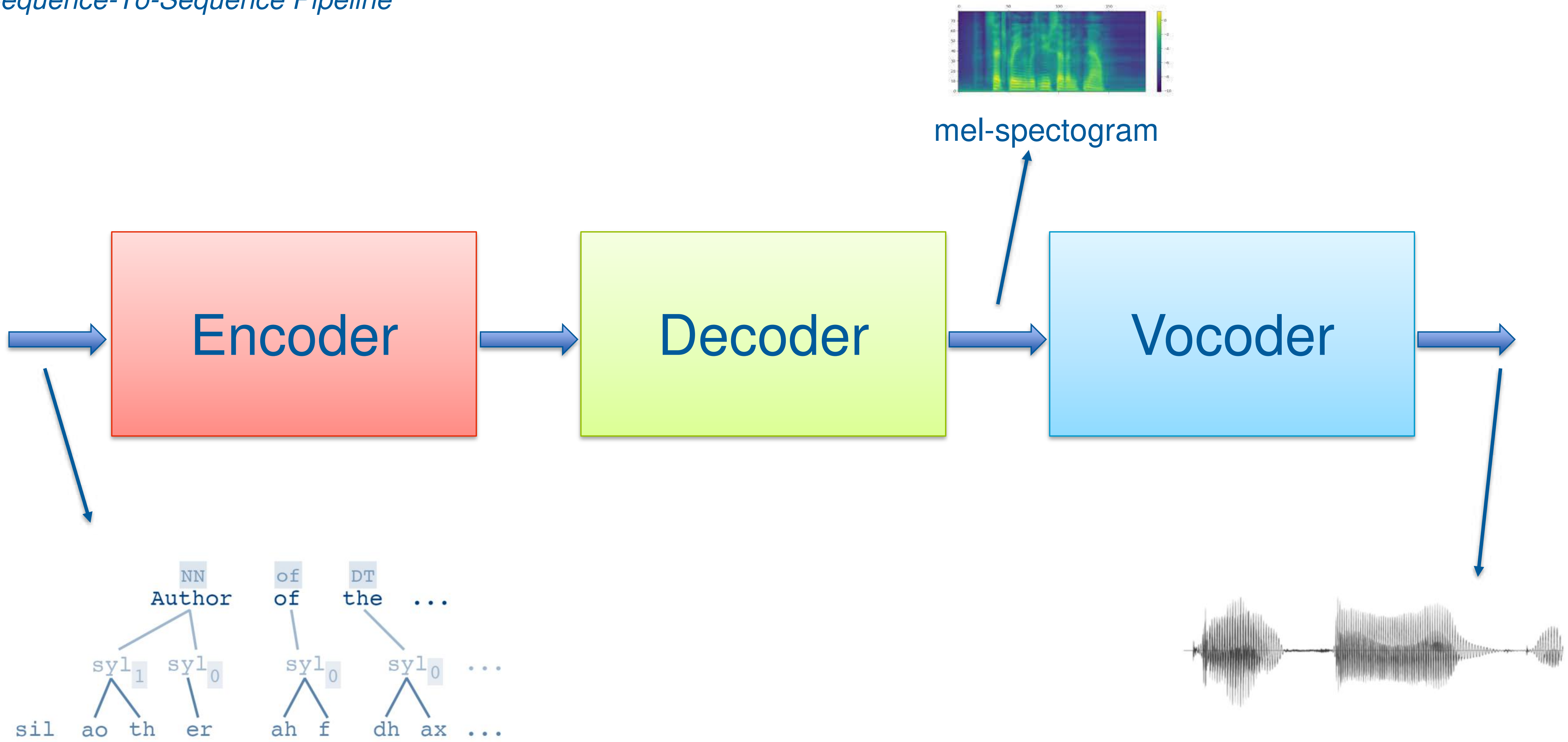
Text-to-Speech Synthesis

Historical Overview



TTS

Sequence-To-Sequence Pipeline



! Active Research Area !

What‘s next?

… in TTS

Active Research areas:

- Neural **Deep learning** approaches
- Better **vocoders** (also neural)
- **Semi- and unsupervised learning.** Why? -> reduce reliance on expensive labelled data
- prosody and its relationship to meaning of text
- Listener- and situation-**appropriate synthesis**

https://speech.zone/courses/speech-synthesis/module-1-introduction/current-technology/
Audio samples from "*PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS*“
Audio samples from "*Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*“

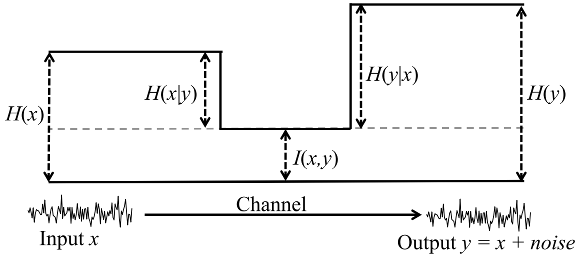
Text-to-Speech Synthesis

Quality measurements

A TTS system is evaluated from different aspects, including intelligibility, naturalness, and preference of the synthetic speech, as well as human perception factors, such as comprehensibility.

Comprehensibility

The degree of received messages being understood

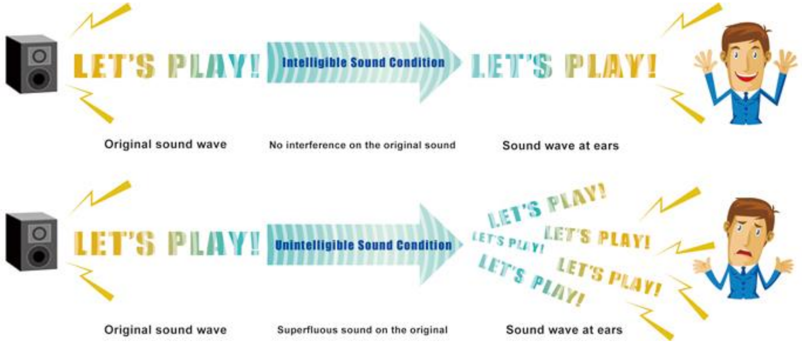


Text-to-Speech Synthesis

Quality measurements

A TTS system is evaluated from different aspects, including intelligibility, naturalness, and preference of the synthetic speech, as well as human perception factors, such as comprehensibility.

Comprehensibility	**Intelligibility**
The degree of received messages being understood	The quality of the audio generated, or the degree of each word being produced in a sentence.



https://www.toa.jp/soundoh/vid/sti/

Text-to-Speech Synthesis

Quality measurements

A TTS system is evaluated from different aspects, including intelligibility, naturalness, and preference of the synthetic speech, as well as human perception factors, such as comprehensibility.

Comprehensibility

The degree of received messages being understood

Intelligibility

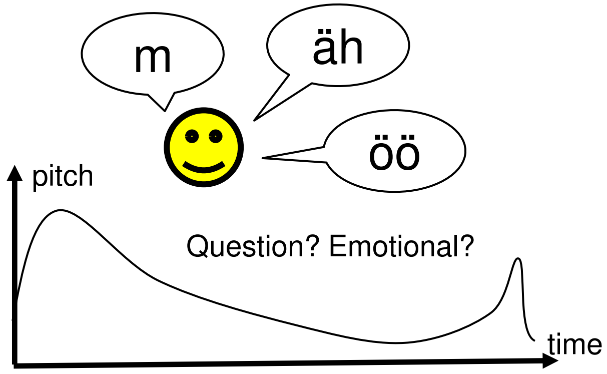
The quality of the audio generated, or the degree of each word being produced in a sentence.

Naturalness

The quality of the speech generated in terms of its timing structure, pronunciation and rendering emotions

"Prosody"?

- intonation (accented syllables; high or low phrase boundaries)
- rhythmic effects (pauses, syllable durations)



Text-to-Speech Synthesis

Quality measurements

A TTS system is evaluated from different aspects, including intelligibility, naturalness, and preference of the synthetic speech, as well as human perception factors, such as comprehensibility.

Comprehensibility

The degree of received messages being understood

Intelligibility

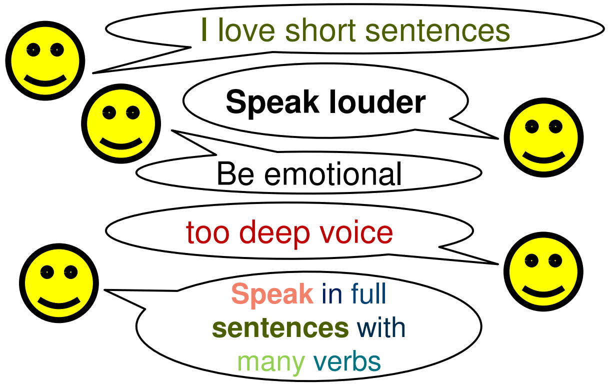
The quality of the audio generated, or the degree of each word being produced in a sentence.

Naturalness

The quality of the speech generated in terms of its timing structure, pronunciation and rendering emotions

Preference

The listeners choice of the better TTS; preference and naturalness are influenced by TTS system, signal quality and voice, in isolation and in combination.



Text-to-Speech Synthesis

Quality measurements

A TTS system is evaluated from different aspects, including intelligibility, naturalness, and preference of the synthetic speech, as well as human perception factors, such as comprehensibility.

Comprehensibility

The degree of received messages being understood

Intelligibility

The quality of the audio generated, or the degree of each word being produced in a sentence.

Naturalness

The quality of the speech generated in terms of its timing structure, pronunciation and rendering emotions

Preference

The listeners choice of the better TTS; preference and naturalness are influenced by TTS system, signal quality and voice, in isolation and in combination.

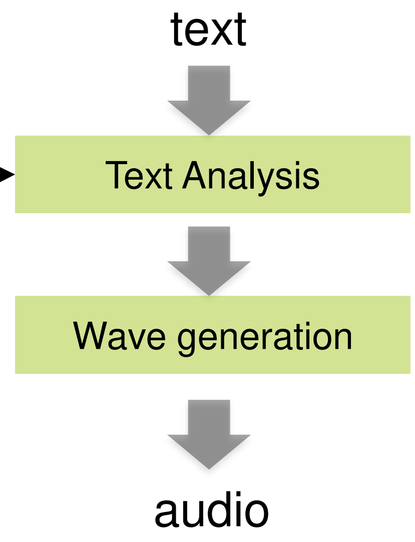
https://www.w3.org/TR/speech-grammar/

Text-to-Speech Synthesis

Preprocessing

A TTS system is evaluated from different aspects, including intelligibility, naturalness, and preference of the synthetic speech, as well as human perception factors, such as comprehensibility.

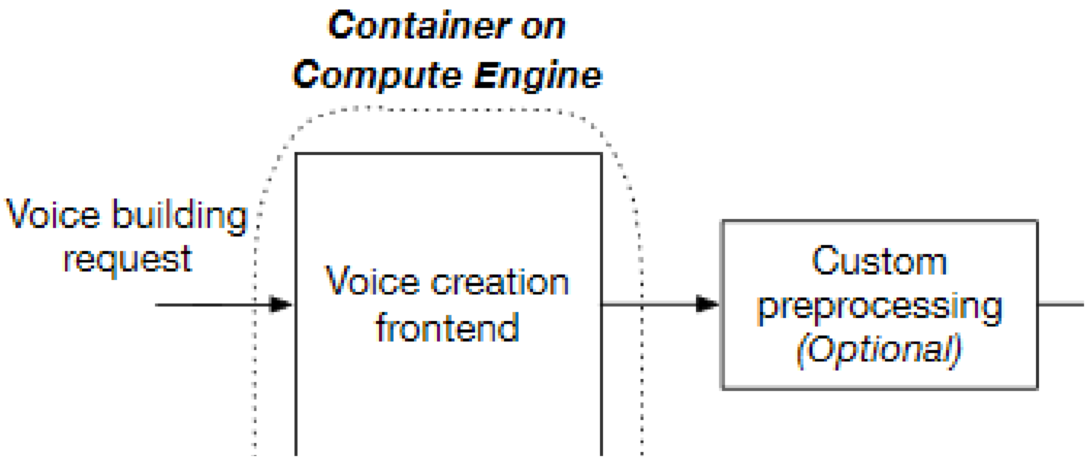
Net patterns (email, web addresses)	Munir.George@THI.De
Date patterns	23/12/2021
Time patterns	10:24 h, 10:24
Duration patterns	11:12 h, 11 h 12 min
Currency patterns	8.95 €
Measure patterns	123.45 km
Telephone number patterns	+49 841 9348-2331
Number patterns (cardinal, ordinal, roman)	23rd III.
Abbreviations	Eng.
Special characters	&



https://ivi.fnwi.uva.nl/cv/events/enterface10/pdf/mary-presentation.pdf

Text-to-Speech Synthesis

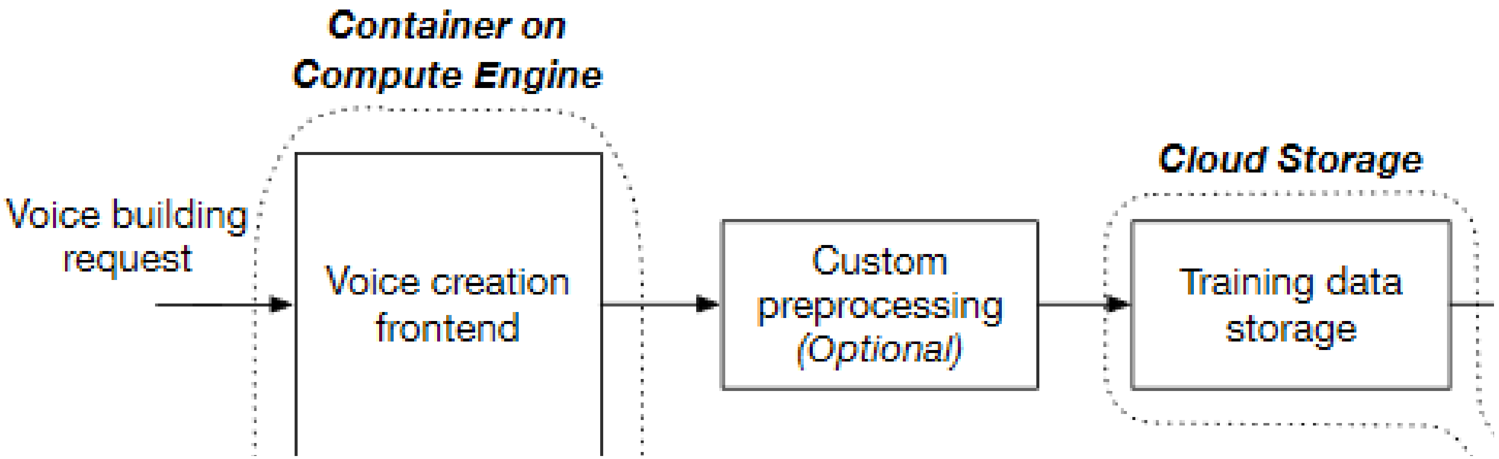
Cloud deployment example


Container on Compute Engine
Voice building request
Voice creation frontend
Custom preprocessing *(Optional)*

https://www.aclweb.org/anthology/L18-1354.pdf

Text-to-Speech Synthesis

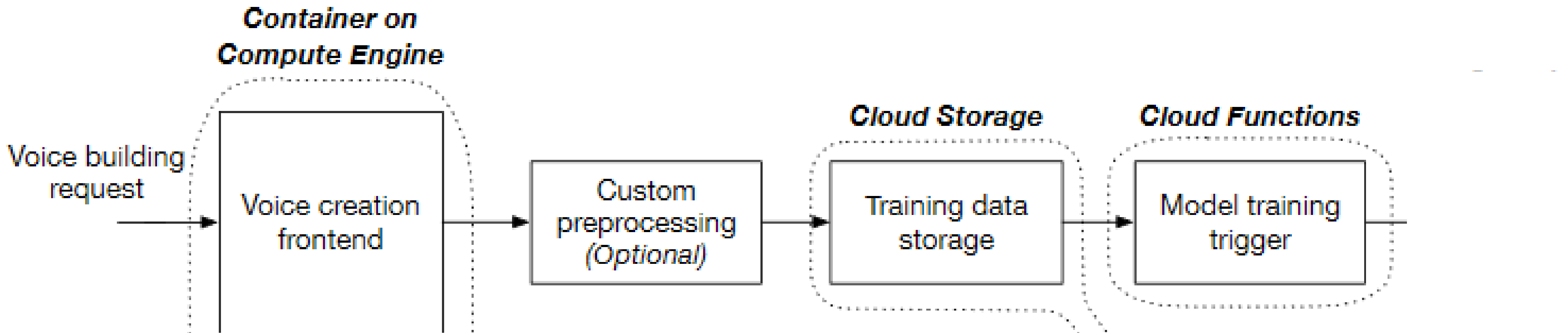
Cloud deployment example



https://www.aclweb.org/anthology/L18-1354.pdf

Text-to-Speech Synthesis

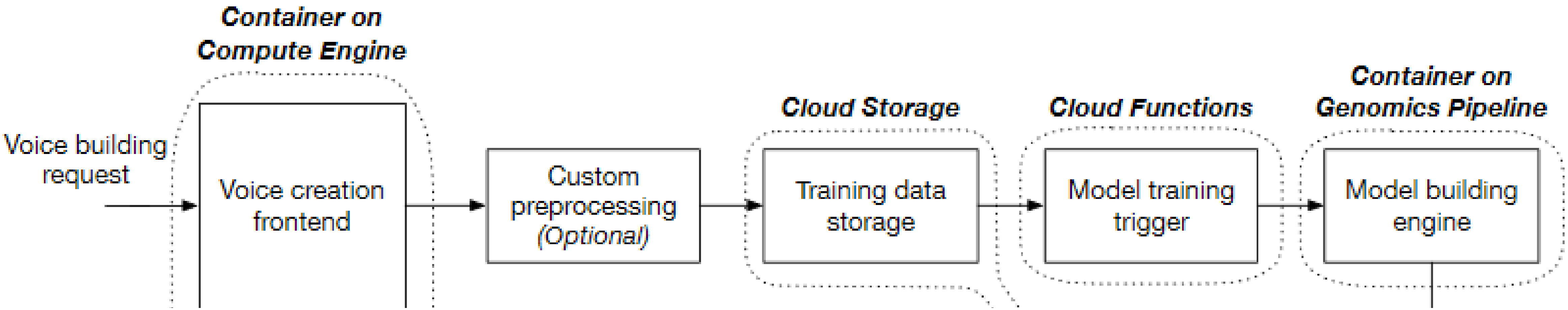
Cloud deployment example



https://www.aclweb.org/anthology/L18-1354.pdf

Text-to-Speech Synthesis

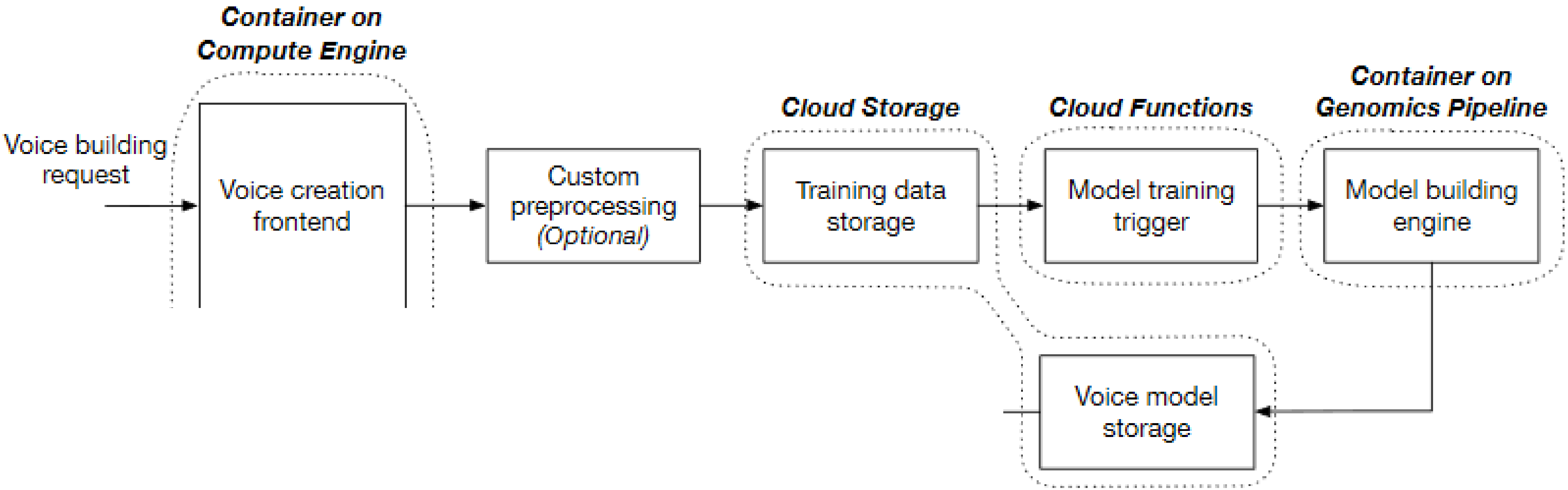
Cloud deployment example



https://www.aclweb.org/anthology/L18-1354.pdf

Text-to-Speech Synthesis

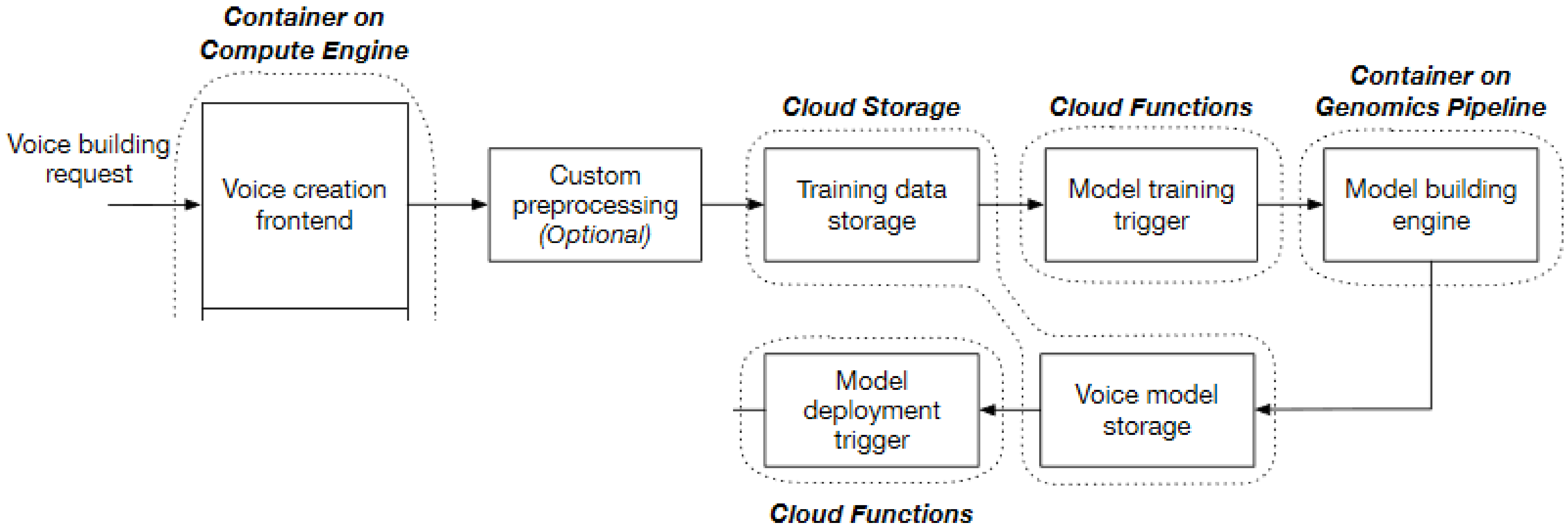
Cloud deployment example



https://www.aclweb.org/anthology/L18-1354.pdf

Text-to-Speech Synthesis

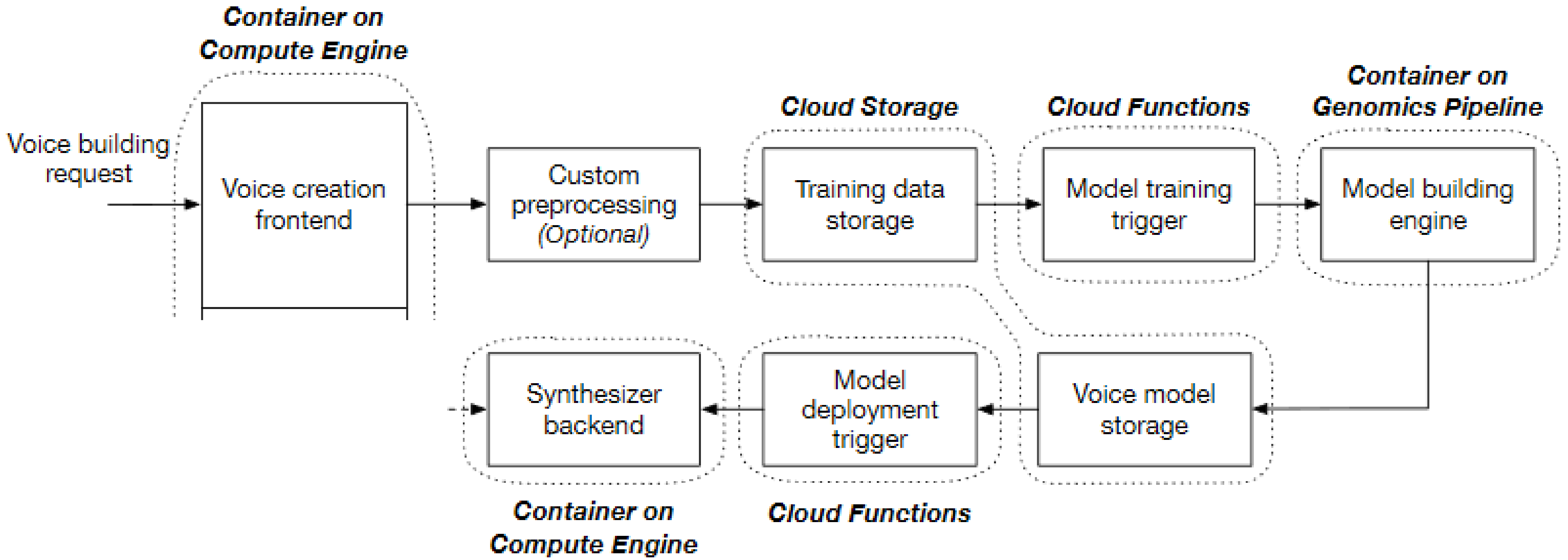
Cloud deployment example



https://www.aclweb.org/anthology/L18-1354.pdf

Text-to-Speech Synthesis

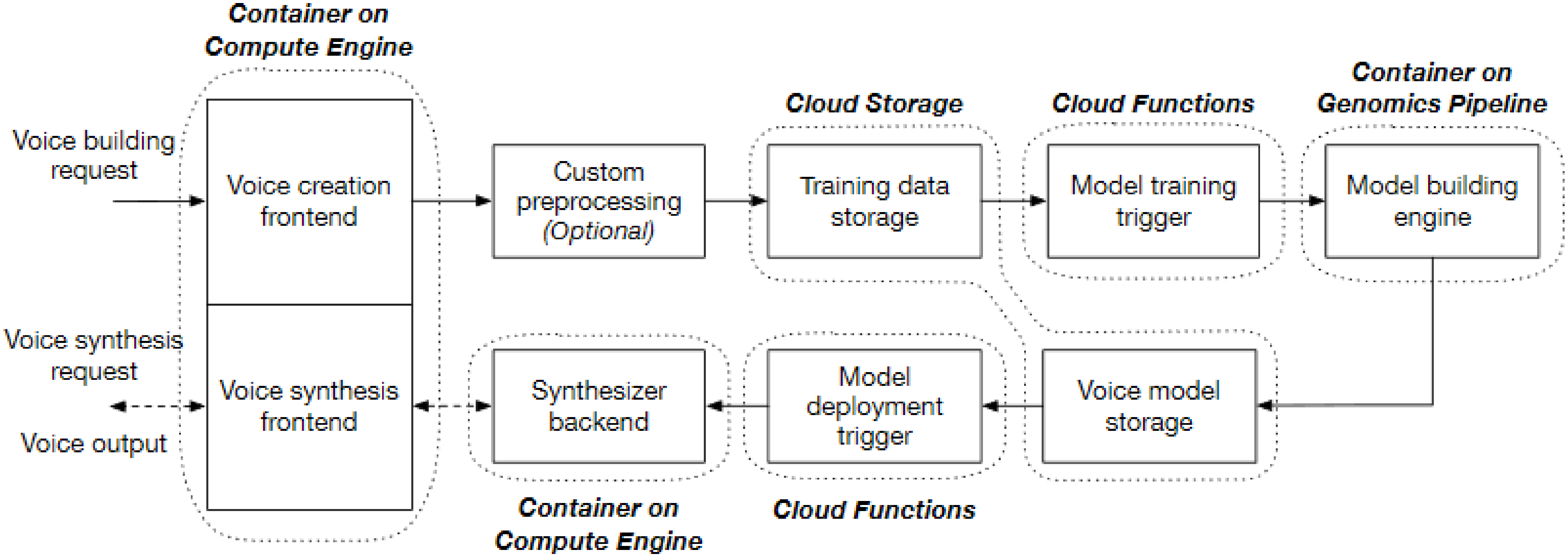
Cloud deployment example



https://www.aclweb.org/anthology/L18-1354.pdf

Text-to-Speech Synthesis

Cloud deployment example



https://www.aclweb.org/anthology/L18-1354.pdf

Along time axis:
https://youtu.be/xzL-pxcpo-E?t=1167

Along frequency axis:
https://youtu.be/xzL-pxcpo-E?t=1184

Text-to-Speech Synthesis

Formant Synthesis

https://learningsynths.ableton.com/

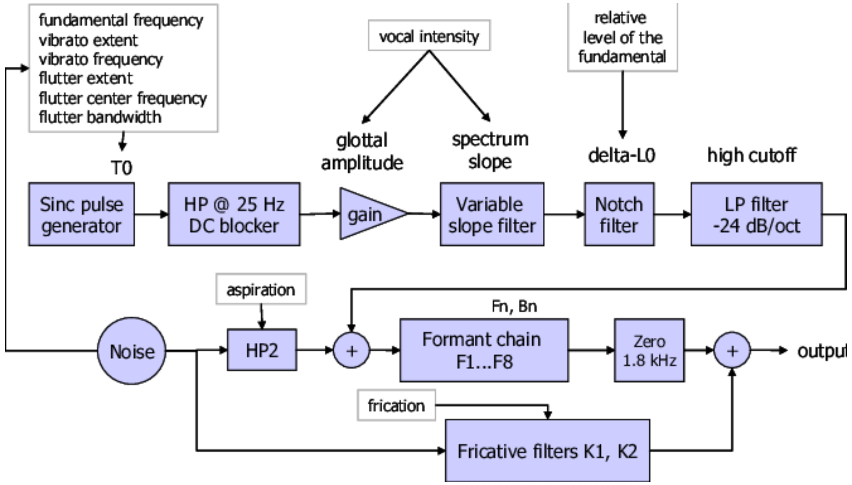
Formants, Spectograms & Vowels (Uni Arizona)
https://synth.playtronica.com/

Text-to-Speech Synthesis

Formant Synthesis

https://learningsynths.ableton.com/

Block diagram of the current KTH formant synthesis model



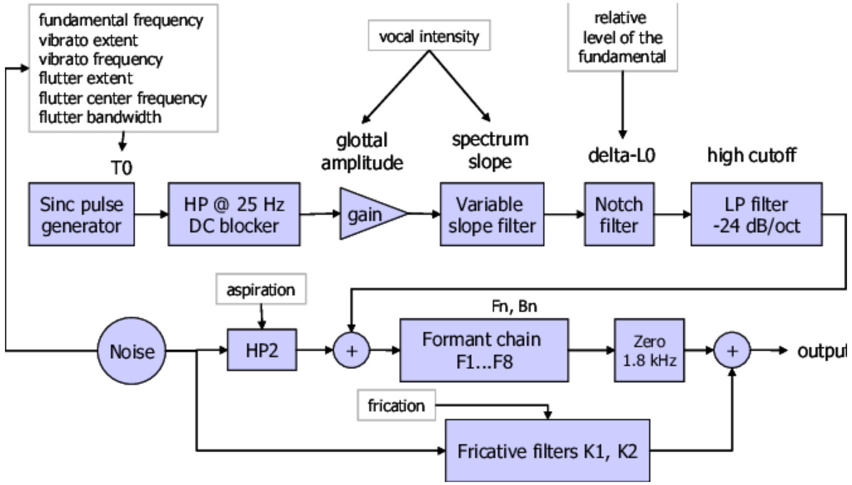
Formants, Spectograms & Vowels (Uni Arizona)
https://synth.playtronica.com/

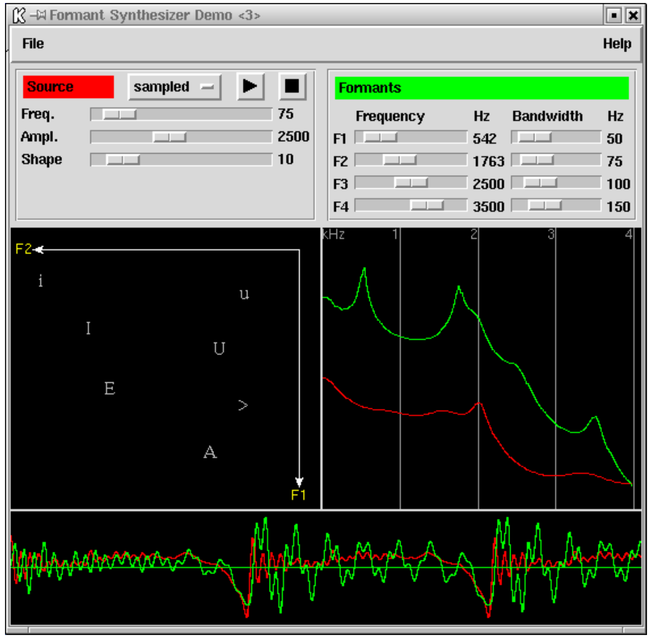
Text-to-Speech Synthesis

Formant Synthesis

https://learningsynths.ableton.com/

Block diagram of the current KTH formant synthesis model





Freq.		75
Ampl.		2500
Shape		10

	Frequency	Hz	Bandwidth	Hz
F1		542		50
F2		1763		75
F3		2500		100
F4		3500		150

Formants, Spectograms & Vowels (Uni Arizona)
https://synth.playtronica.com/

Text-to-Speech Synthesis

Formant Synthesis

- Rule-based TTS produces speech segments by generating artificial signals based on a set of specified rules **mimicking the formant structure** and other **spectral properties** of natural speech.

Text-to-Speech Synthesis

Formant Synthesis

- Rule-based TTS produces speech segments by generating artificial signals based on a set of specified rules **mimicking the formant structure** and other **spectral properties** of natural speech.
- The synthesized speech is produced using an **additive synthesis** and an **acoustic model**.

Text-to-Speech Synthesis

Formant Synthesis

- Rule-based TTS produces speech segments by generating artificial signals based on a set of specified rules **mimicking the formant structure** and other **spectral properties** of natural speech.
- The synthesized speech is produced using an **additive synthesis** and an **acoustic model**.
- The acoustic model uses parameters like, **voicing, fundamental frequency, noise levels, etc**. that varied over time.

Text-to-Speech Synthesis

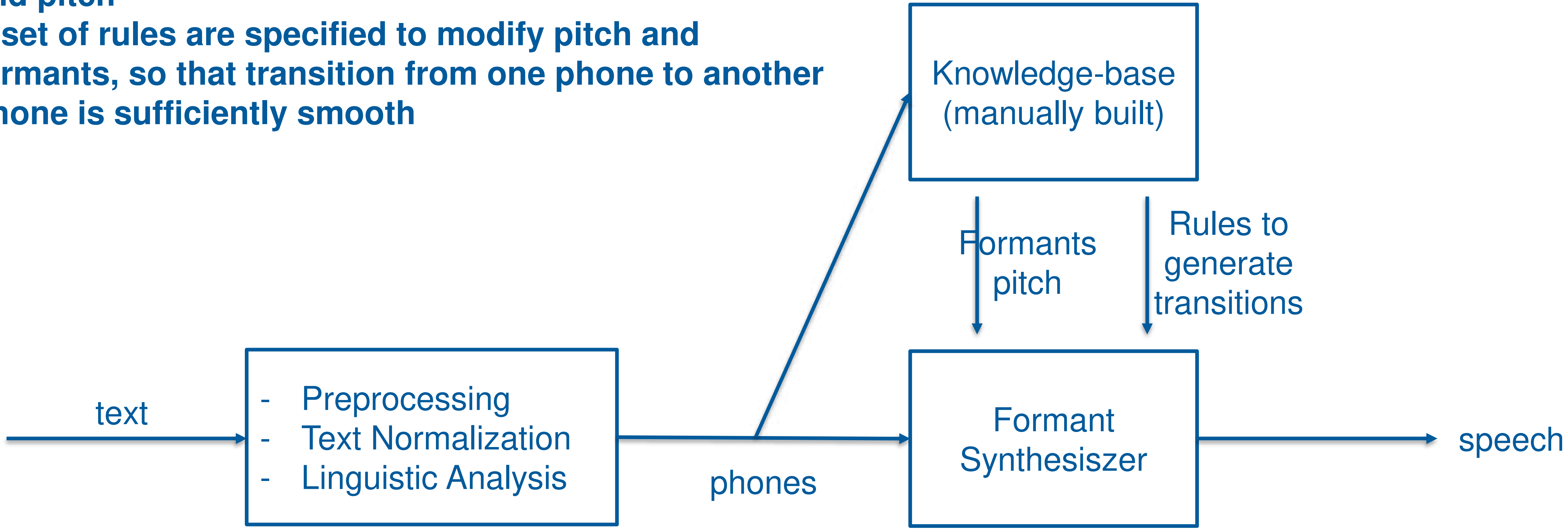
Formant Synthesis

- Rule-based TTS produces speech segments by generating artificial signals based on a set of specified rules **mimicking the formant structure** and other **spectral properties** of natural speech.
- The synthesized speech is produced using an **additive synthesis** and an **acoustic model**.
- The acoustic model uses parameters like, **voicing, fundamental frequency, noise levels, etc**. that varied over time.
- Formant-based systems can **control all aspects of the output speech**, producing a wide variety of emotions and different tone voices.

Formant Synthesis

Architecture

- **Each phone is produced by specifying the formants and pitch**
- **A set of rules are specified to modify pitch and formants, so that transition from one phone to another phone is sufficiently smooth**



Text-to-Speech Synthesis

Formant Synthesis

Rule-based TTS produces speech segments by generating artificial signals based on a set of specified rules mimicking the formant structure and other spectral properties of natural speech. The synthesized speech is produced using an additive synthesis and an acoustic model. The acoustic model uses parameters like, voicing, fundamental frequency, noise levels, etc. that varied over time. Formant-based systems can control all aspects of the output speech, producing a wide variety of emotions and different tone voices.

Good things:

- Highly intelligible synthesized speech, even at high speeds, avoiding the acoustic glitches.
- Less dependent on a speech corpus to produce the output speech.
- Well-suited for embedded systems, where memory and microprocessor power are limited.

Text-to-Speech Synthesis

Formant Synthesis

Rule-based TTS produces speech segments by generating artificial signals based on a set of specified rules mimicking the formant structure and other spectral properties of natural speech. The synthesized speech is produced using an additive synthesis and an acoustic model. The acoustic model uses parameters like, voicing, fundamental frequency, noise levels, etc. that varied over time. Formant-based systems can control all aspects of the output speech, producing a wide variety of emotions and different tone voices.

Good things:

- Highly intelligible synthesized speech, even at high speeds, avoiding the acoustic glitches.
- Less dependent on a speech corpus to produce the output speech.
- Well-suited for embedded systems, where memory and microprocessor power are limited.

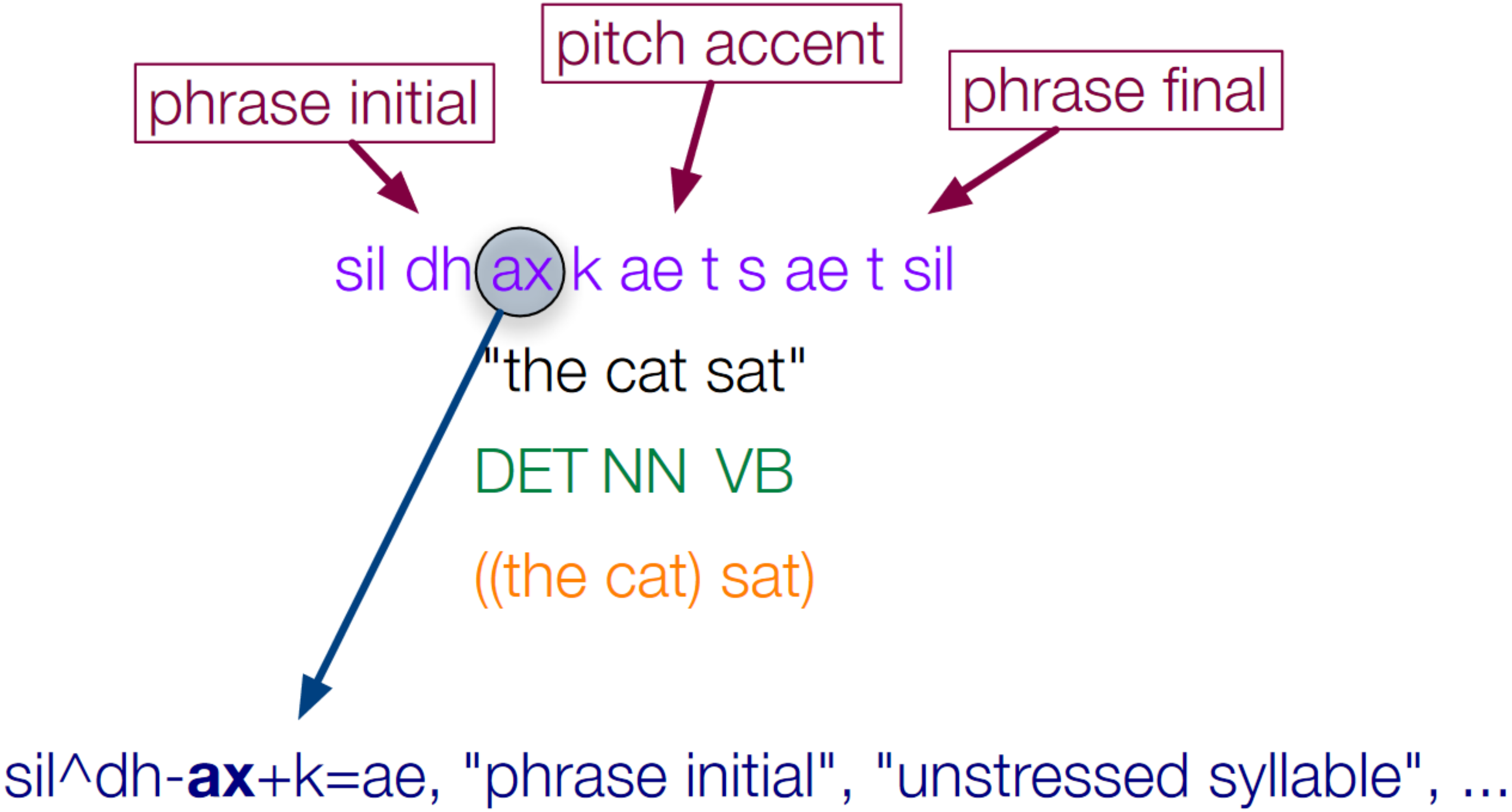
Bad things:

- Low naturalness: the technique produces artificial, robotic-sounding speech that is far from the natural speech spoken by a human.
- Difficult to design rules that specify the timing of the source and the dynamic values of all filter parameters for even simple words

Formants, Spectograms & Vowels (Uni Arizona)
https://synth.playtronica.com/

Text-to-Speech

Overview



phrase initial

pitch accent

phrase final

sil dh ax k ae t s ae t sil

"the cat sat"

DET NN VB

((the cat) sat)

sil^dh-**ax**+k=ae, "phrase initial", "unstressed syllable", ...

http://mi.eng.cam.ac.uk/foswiki/pub/Main/SeminarsSpeech/Cam-SpeechSynthesis-Seminar.pdf

Text-to-Speech Synthesis

Concatenative TTS

General idea: Use pre-recorded speech units to generate new speech

Text-to-Speech Synthesis

Concatenative TTS

General idea: Use pre-recorded speech units to generate new speech

- Voice actors are recorded saying a range of speech units, (sentences, syllables).

Text-to-Speech Synthesis

Concatenative TTS

General idea: Use pre-recorded speech units to generate new speech

- Voice actors are recorded saying a range of speech units, (sentences, syllables).

- These are further labeled and segmented by linguistic units (phones, phrases, sentences) forming a huge database.

Text-to-Speech Synthesis

Concatenative TTS

General idea: Use pre-recorded speech units to generate new speech

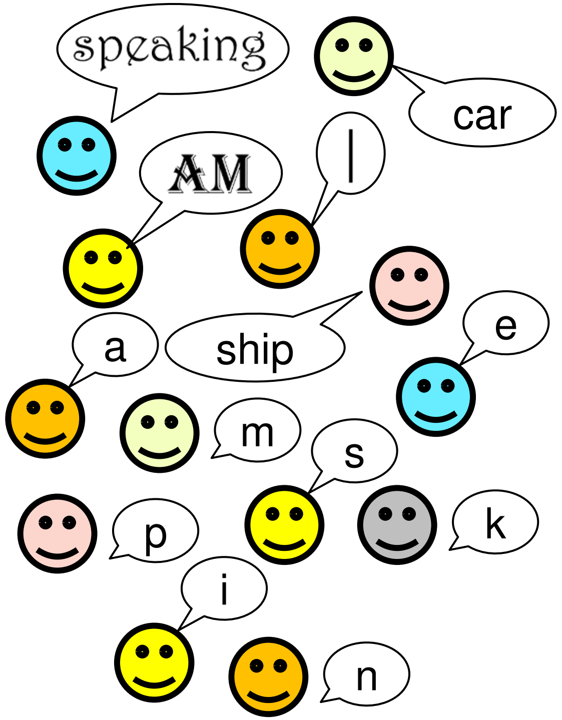
- Voice actors are recorded saying a range of speech units, (sentences, syllables).
- These are further labeled and segmented by linguistic units (phones, phrases, sentences) forming a huge database.
- During speech synthesis, a Text-to-Speech engine searches such database for speech units that match the input text, concatenates them together and produces an audio file.

Text-to-Speech Synthesis

Concatenative TTS

General idea: Use pre-recorded speech units to generate new speech

Recordings of high-quality audio clips are combined to form the speech. Voice actors are recorded saying a range of speech units, (sentences, syllables). These are further labeled and segmented by linguistic units (phones, phrases, sentences) forming a huge database. During speech synthesis, a Text-to-Speech engine searches such database for speech units that match the input text, concatenates them together and produces an audio file.

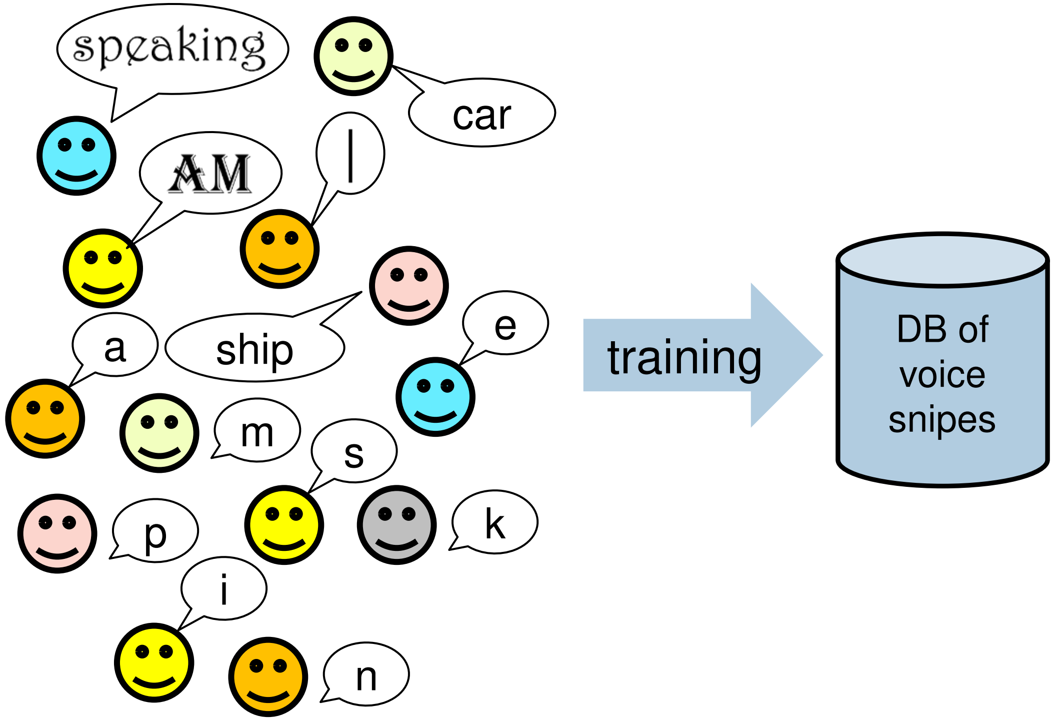


Text-to-Speech Synthesis

Concatenative TTS

General idea: Use pre-recorded speech units to generate new speech

Recordings of high-quality audio clips are combined to form the speech. Voice actors are recorded saying a range of speech units, (sentences, syllables). These are further labeled and segmented by linguistic units (phones, phrases, sentences) forming a huge database. During speech synthesis, a Text-to-Speech engine searches such database for speech units that match the input text, concatenates them together and produces an audio file.

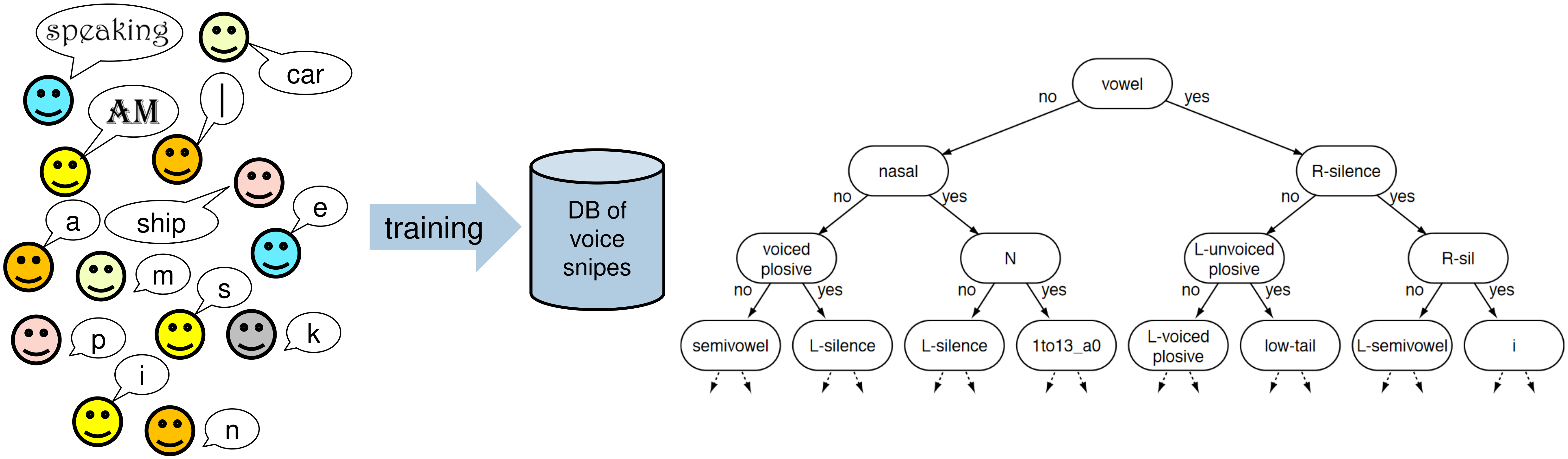


Text-to-Speech Synthesis

Concatenative TTS

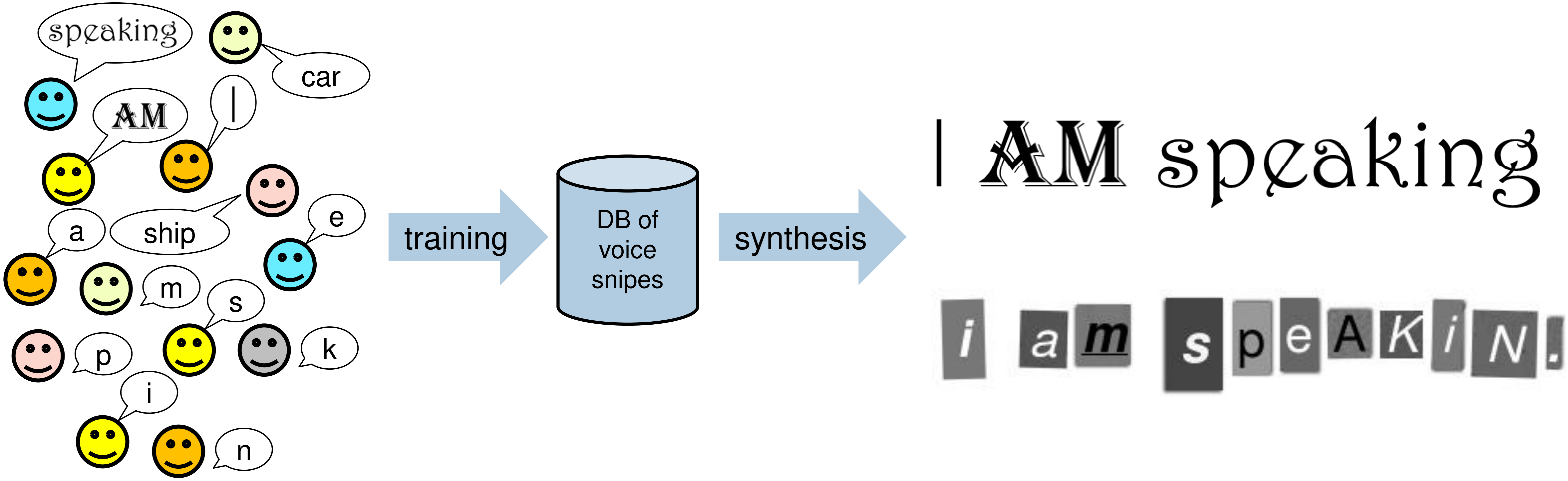
General idea: Use pre-recorded speech units to generate new speech

Recordings of high-quality audio clips are combined to form the speech. Voice actors are recorded saying a range of speech units, (sentences, syllables). These are further labeled and segmented by linguistic units (phones, phrases, sentences) forming a huge database. During speech synthesis, a Text-to-Speech engine searches such database for speech units that match the input text, concatenates them together and produces an audio file.



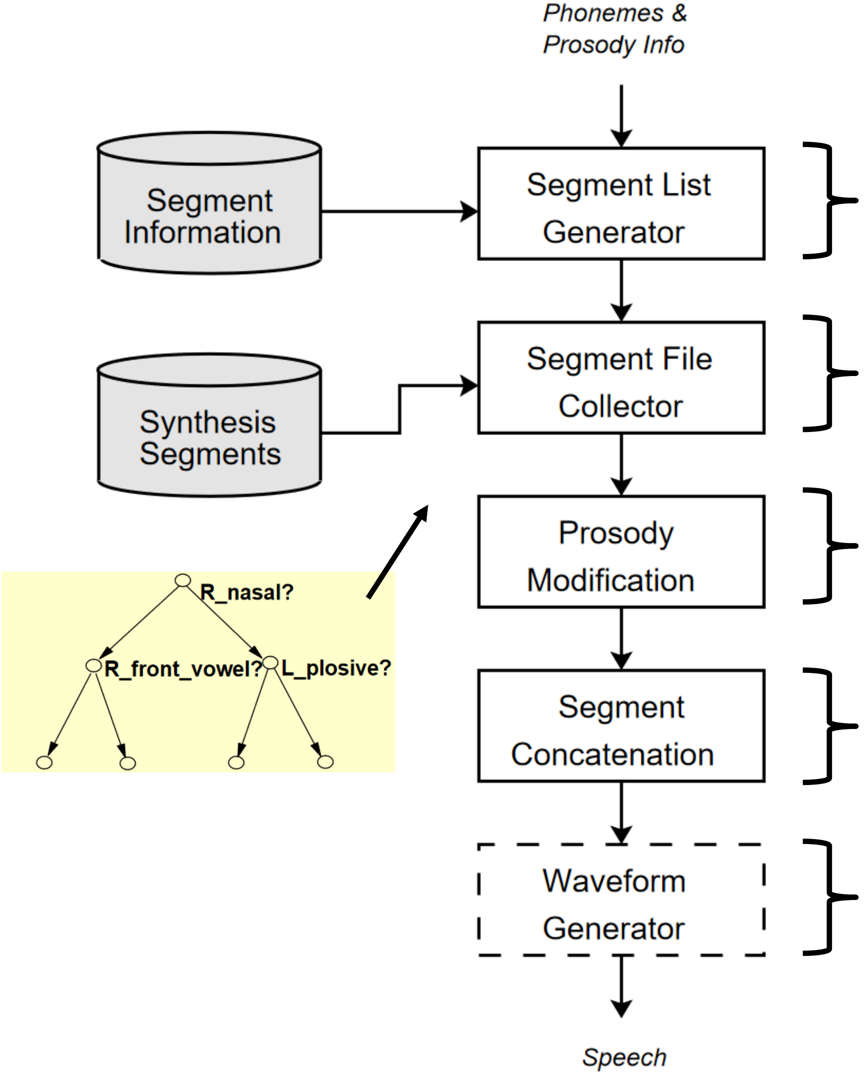
General idea: Use pre-recorded speech units to generate new speech

Recordings of high-quality audio clips are combined to form the speech. Voice actors are recorded saying a range of speech units, (sentences, syllables). These are further labeled and segmented by linguistic units (phones, phrases, sentences) forming a huge database. During speech synthesis, a Text-to-Speech engine searches such database for speech units that match the input text, concatenates them together and produces an audio file.



Text-to-Speech Synthesis

Concatenative TTS



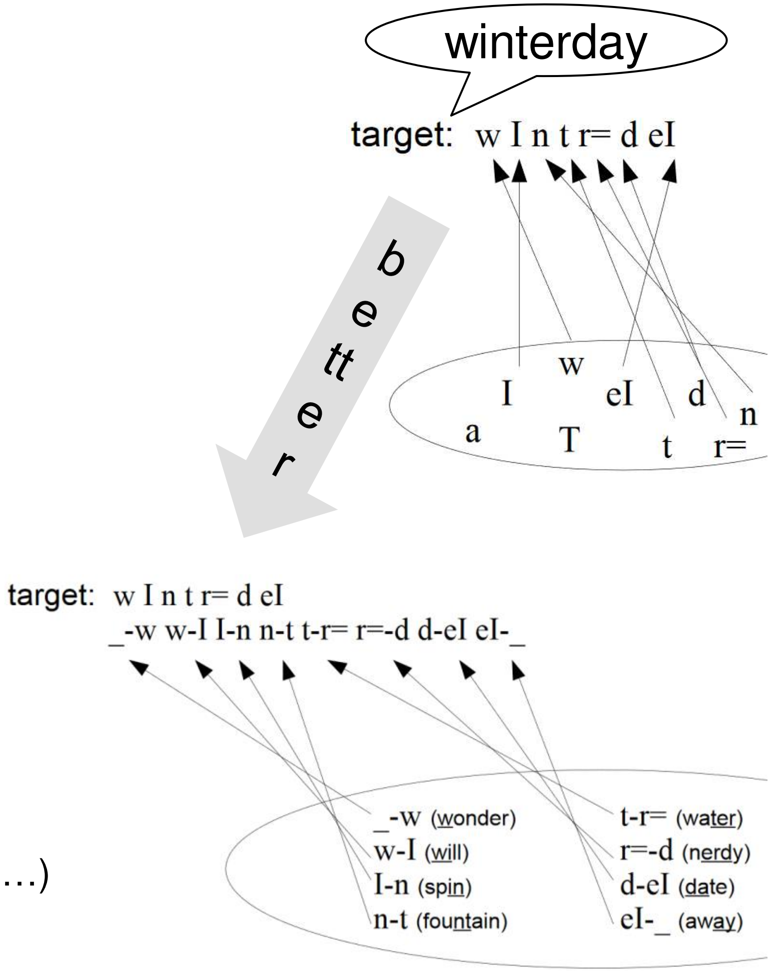
Phonetic input notation is transformed into the pre-defined segment format of the synthesizer.

Maps the phonetic segments with the synthesis segments

Adds prosodic properties to the speech segments (pitch, time duration, etc.)

Concatenation of speech segments

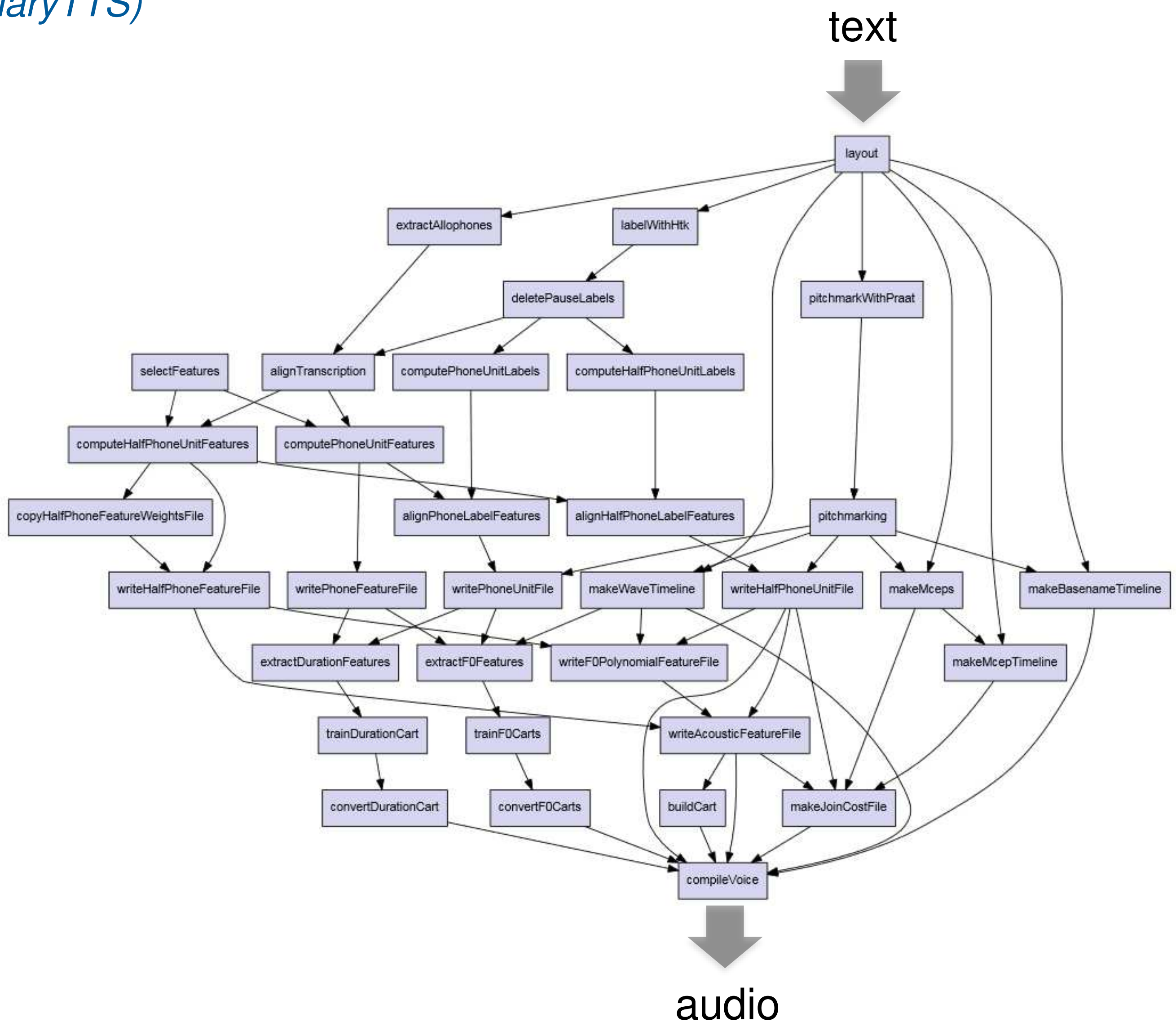
Generates the waveform in desired format (mp3, …)



http://www.cs.columbia.edu/~ecooper/tts/SS_Lecture_CUNY_noaudio.pdf
https://www.diva-portal.org/smash/get/diva2:1022952/FULLTEXT01.pdf

Text-to-Speech Synthesis

Concatenative TTS (MaryTTS)



http://www.coli.uni-saarland.de/~steiner/teaching/2014/winter/voicebuilding/intro.pdf
https://ivi.fnwi.uva.nl/cv/events/enterface10/pdf/mary-presentation.pdf

Text-to-Speech Synthesis

Concatenative TTS: (Dis-)Advantages

Recordings of high-quality audio clips are combined to form the speech. Voice actors are recorded saying a range of speech units, (sentences, syllables). These are further labeled and segmented by linguistic units (phones, phrases, sentences) forming a huge database. During speech synthesis, a Text-to-Speech engine searches such database for speech units that match the input text, concatenates them together and produces an audio file.

Good things:

- High quality of audio in terms of intelligibility;
- Possibility to preserve the original actor’s voice;

Text-to-Speech Synthesis

Concatenative TTS: (Dis-)Advantages

Recordings of high-quality audio clips are combined to form the speech. Voice actors are recorded saying a range of speech units, (sentences, syllables). These are further labeled and segmented by linguistic units (phones, phrases, sentences) forming a huge database. During speech synthesis, a Text-to-Speech engine searches such database for speech units that match the input text, concatenates them together and produces an audio file.

Good things:

- High quality of audio in terms of intelligibility;
- Possibility to preserve the original actor's voice;

Bad things:

- Such systems are very time consuming because they require huge databases, and hard-coding the combination to form these words;
- The resulting speech may sound less natural and emotionless, because it is nearly impossible to get the audio recordings of all possible words spoken in all possible combinations of emotions, prosody, stress, etc.