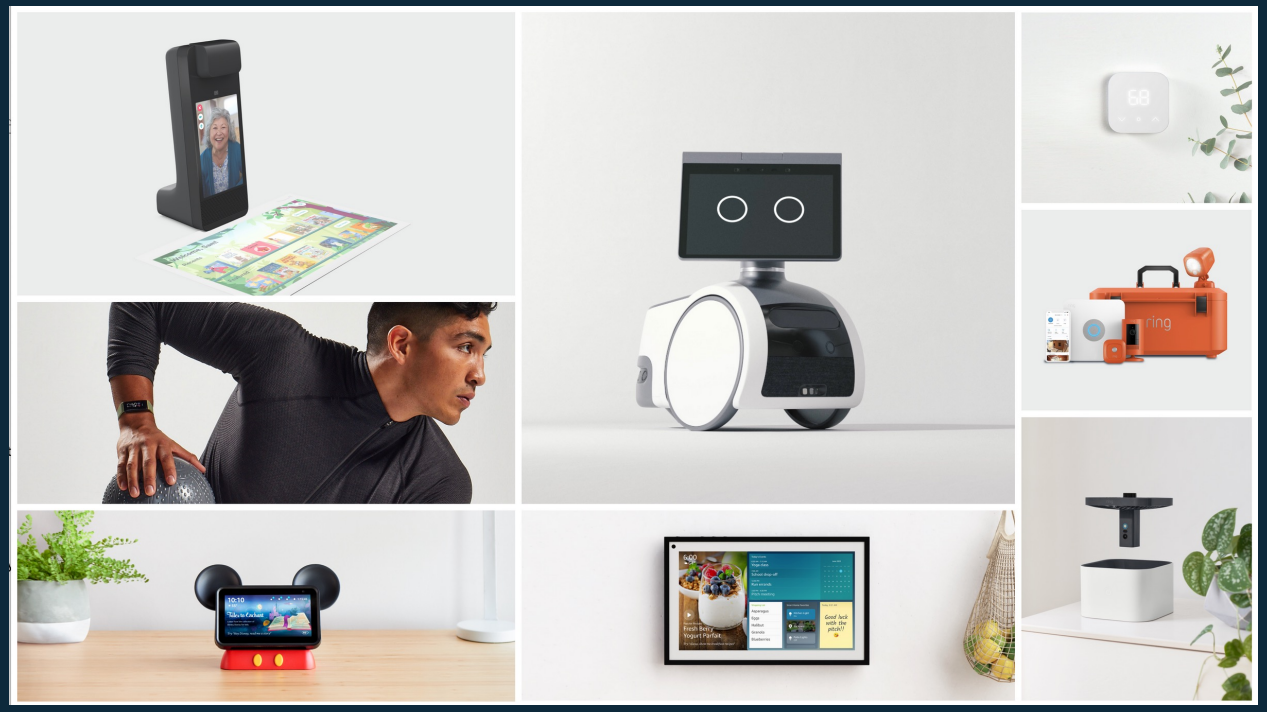
Expressive Neural Text-to-Speech

Andrew Breen

The Voice of Amazon Products



Available in 16 languages

Amazon Polly

Available in 63 voices in 29 languages

Alexa

Utterance

Alexa Voice Service

Recognize

Automatic Speech Recognition (ASR)

Recognition result

"Play" directive

Recognition result

Natural Language Understanding (NLU)

Intent

Text to Speech (TTS)

Intent

Skills

"Play" directive

Brief History of Text-to-Speech (TTS)

1939

The VODER

The first electronic voice Synthesizer

1980

DECTALK COMMERCIAL

Formant based TTS system

1990

DIPHONE

Speech synthesis based on human recordings

2000

CONCATENATIVE TTS

First commercial concatenative TTS systems appear

2014

CONCATENATIVE TTS

Echo released with Alexa Concatenative TTS

2017

CONCATENATIVE HYBRID TTS

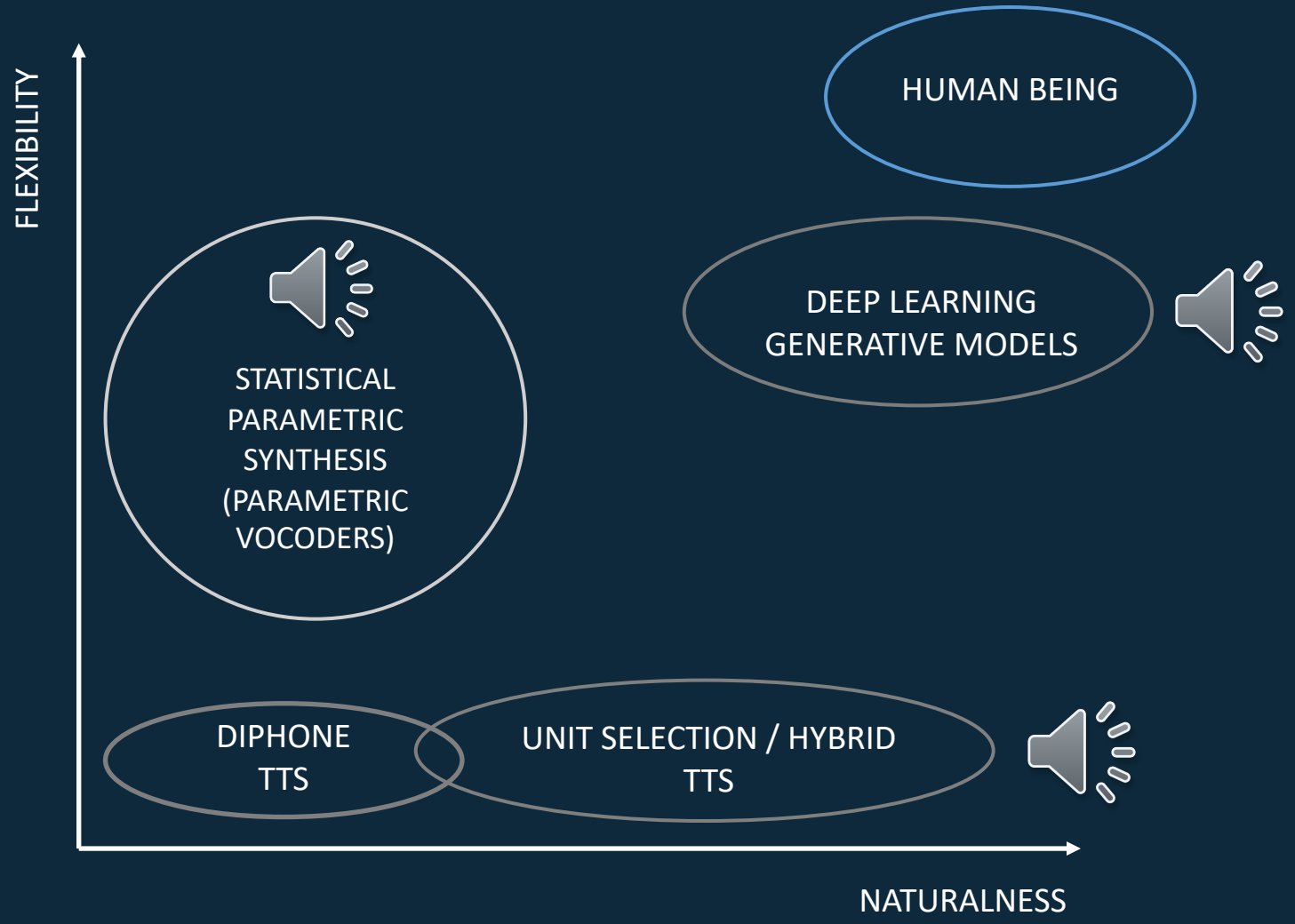
Latest Alexa concatenative

2018

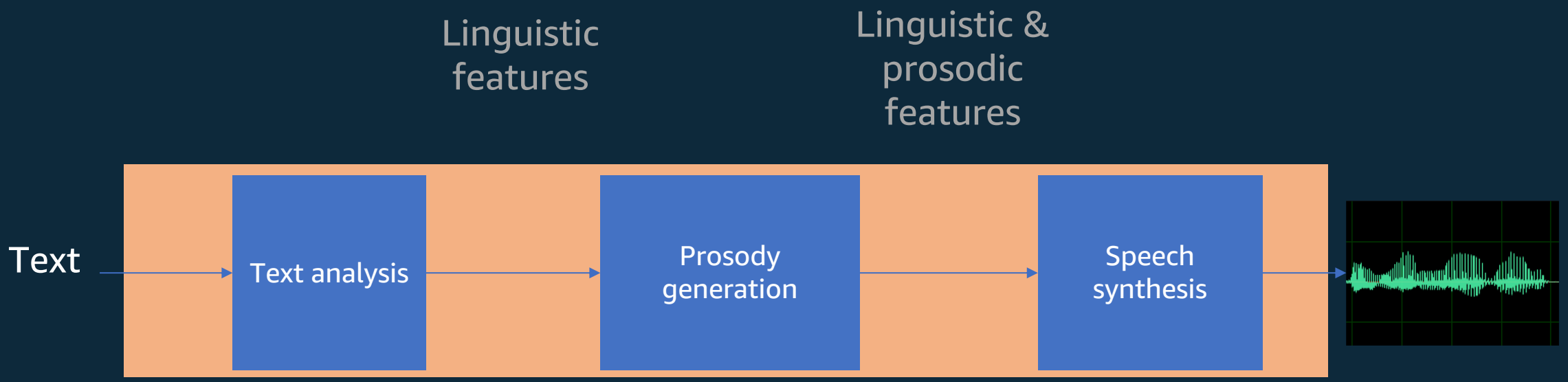
NEURAL TTS

First applications using Alexa Neural TTS

Comparison of TTS technologies

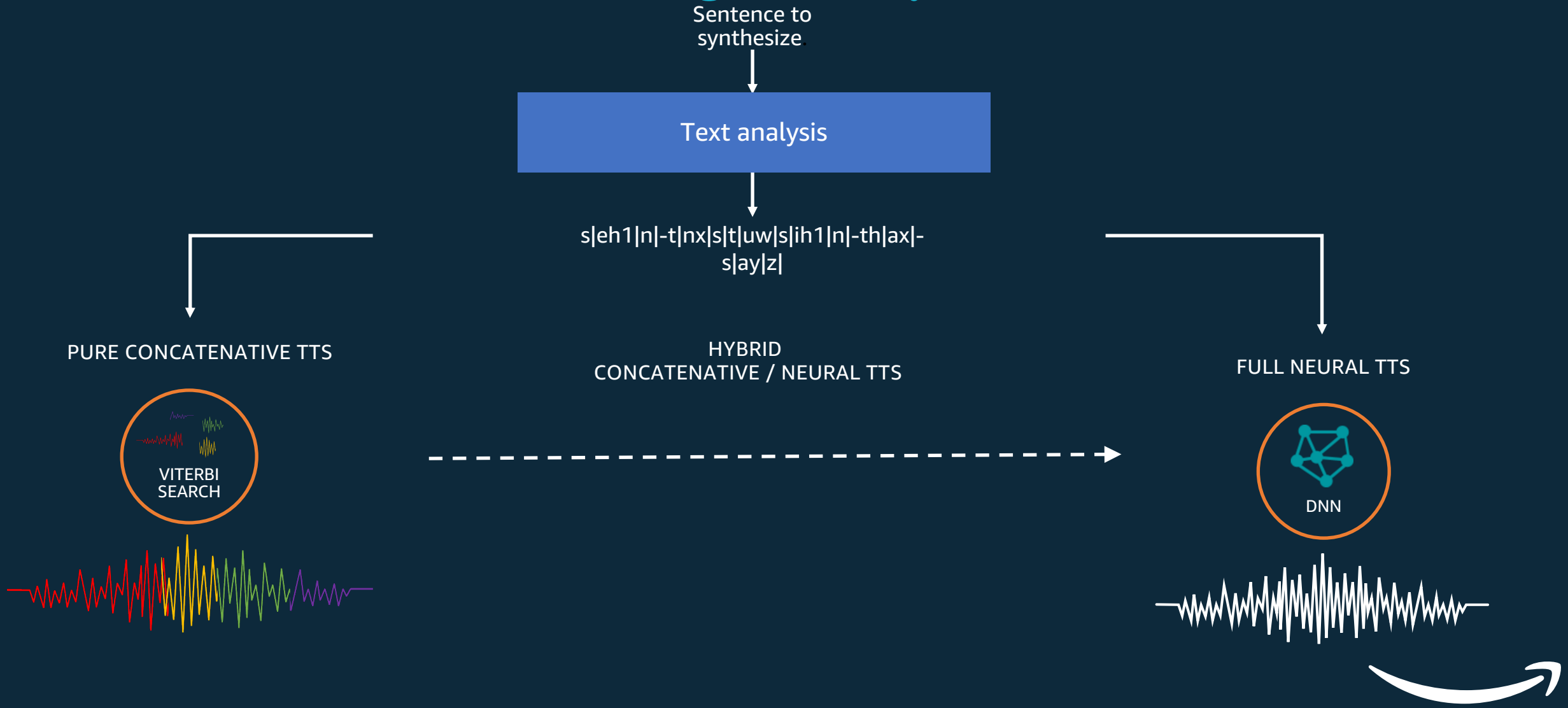


Components of a TTS system

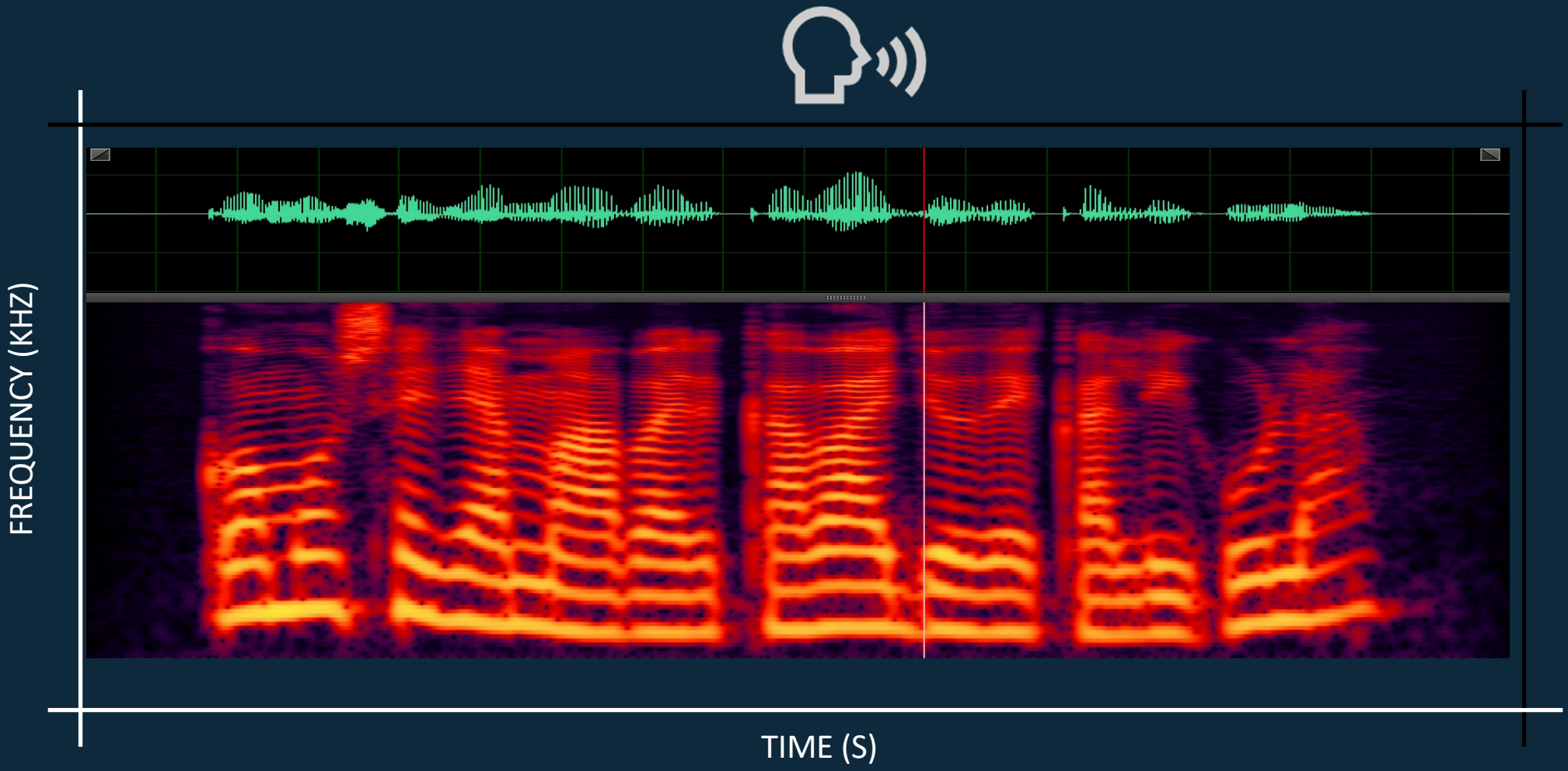


- Extract linguistic knowledge
- Syntax, part of speech, sounds of language
- Mathematical models of style
- Speech rate, pitch, & phrasing
- Generation of acoustic signal

Neural text-to-speech (NTTS) brings together natural prosody & high fidelity



How Speech is Represented



Neural Text-to-Speech

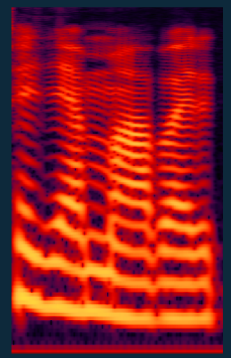
PHONEMES &
LINGUISTIC FEATURES

ACOUSTIC MODELING

(Sequence-to-sequence network)

- Focused on speech representation
- Speaker and style specific

MEL
SPECTROGRAMS



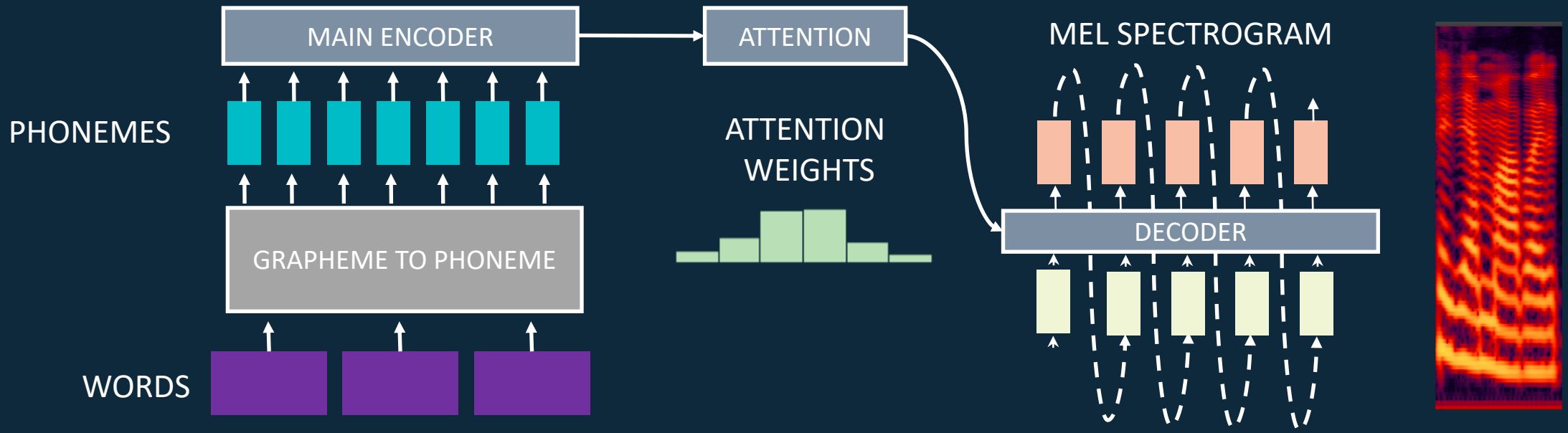
AUDIO SYNTHESIS

(Universal Neural Vocoder)

- Focused on Hi-Fi speech signal
- Independent of speaker or style

SPEECH

Acoustic modeling



Alexa NTTS

Natural

Neural Text-to-Speech

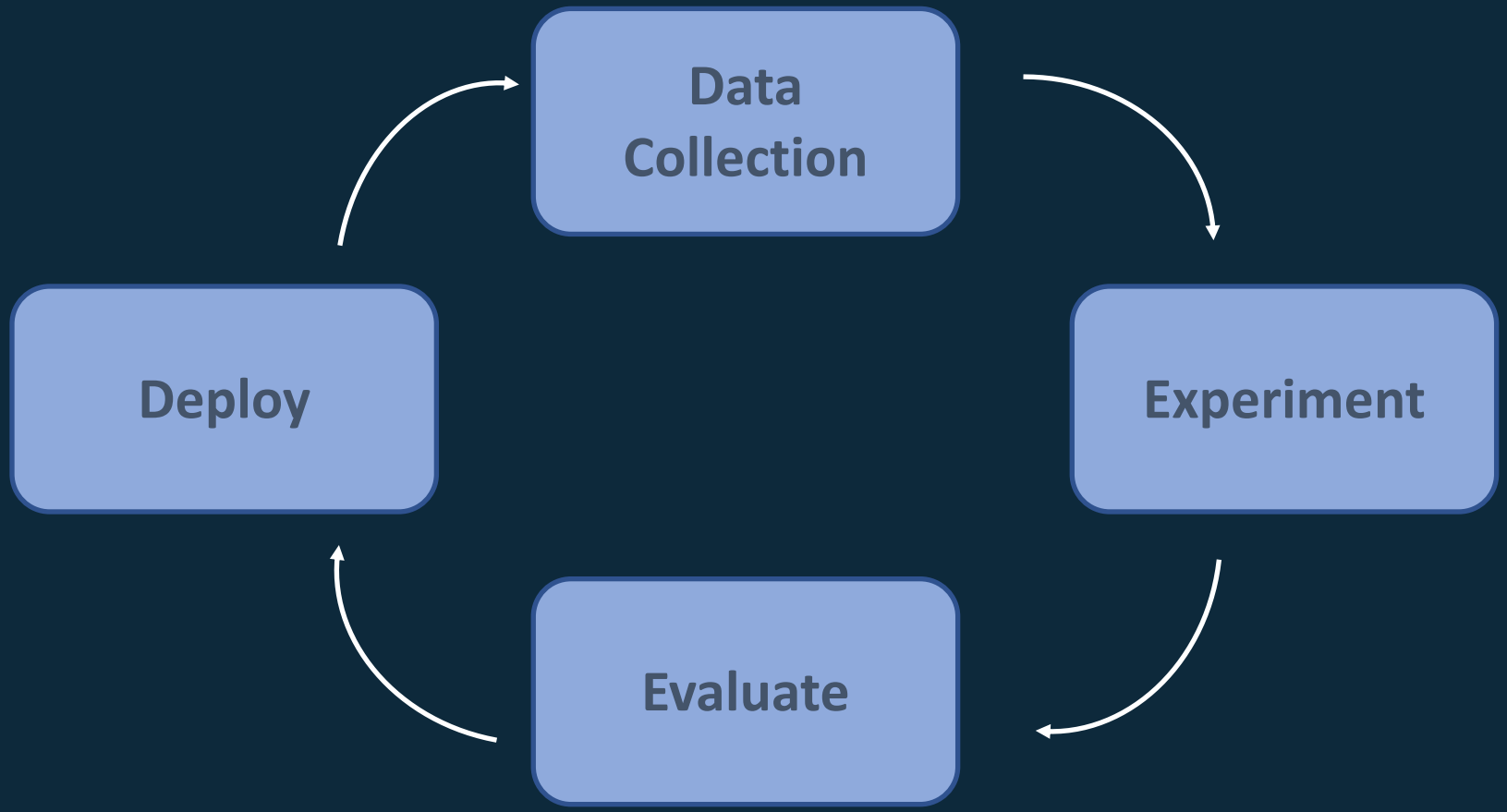
Introducing Amazon's first custom electric delivery vehicle

“Amazon has revealed its first custom electric delivery vehicle, designed and built in partnership with Rivian, and expects to have 10,000 of the new vans on the road delivering to customers as early as 2022.”

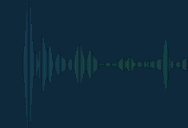
Unit Selection

Neural TTS

The TTS development cycle

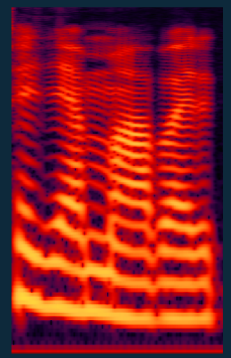


Neural Text-to-Speech



SPEECH

MEL
SPECTROGRAMS

ACOUSTIC MODELING →  → AUDIO SYNTHESIS

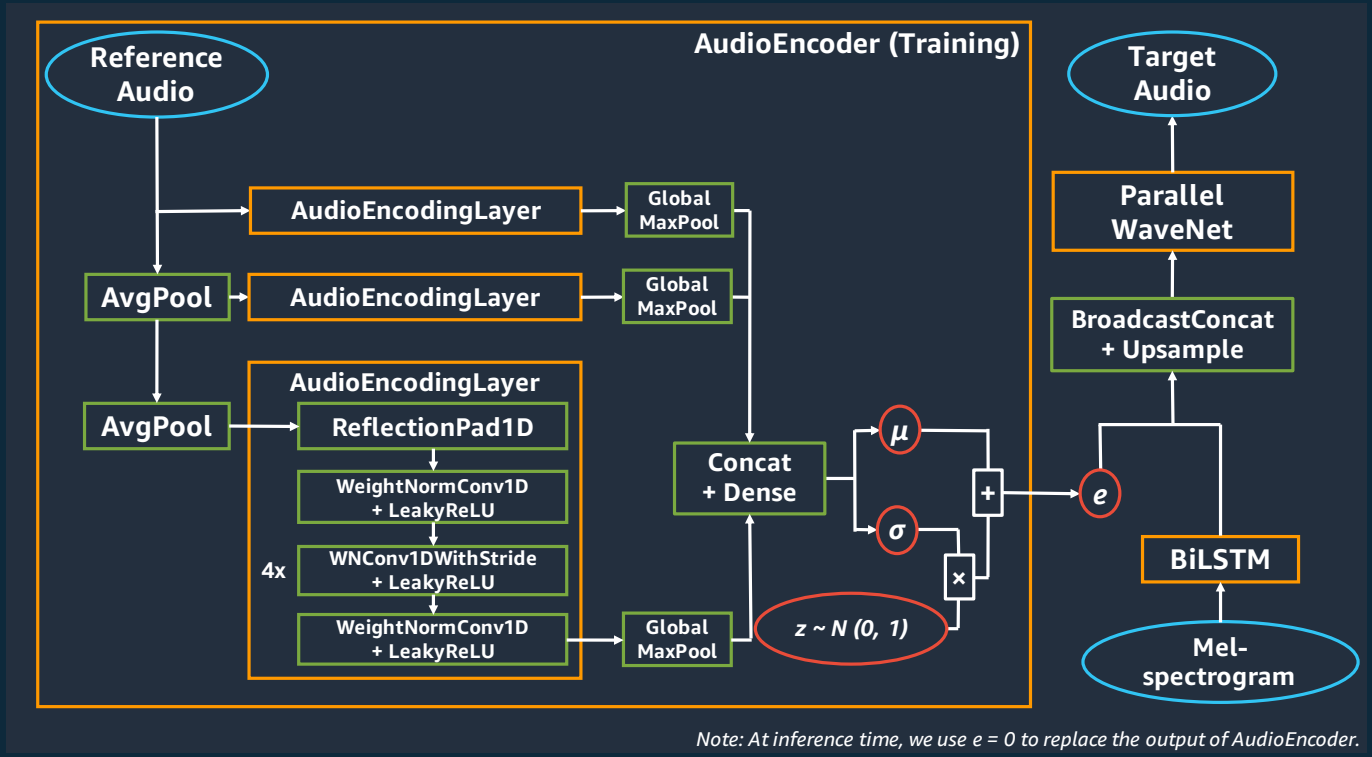
(Universal Neural Vocoder)

- Focused on Hi-Fi speech signal
- Independent of speaker or style

Universal Neural Vocoding

Universal Parallel WaveNet (UPW)

- We trained a **universal neural vocoder** based on Parallel WaveNet, using a multi-speaker multi-lingual high-quality speech corpus.
- In order to train a universal vocoder, we propose an additional VAE-type conditioning network called **Audio Encoder**.



Recording

Speaker Dedicated PW

Baseline PW on multi-speaker data

Proposed Universal PW

Y. Jiao et al., "Universal Neural Vocoding with Parallel WaveNet", ICASSP 2021

Universal Neural Vocoding

Comparison with speaker-dependent vocoders

MUSHRA results per voice

MUSHRA	Recording	SDPW	UPW
All internal	69.68	57.92	**58.70**
British Eng. / F / Adult	71.64	65.69	**67.67**
Aus. Eng. / M / Adult	73.52	**68.37**	68.32
Spanish / F / Adult	69.06	60.27	**61.17**
Indian Eng. / F / Adult	77.19	62.22	**66.95**
*US Eng. / M / Senior	70.40	57.65	**60.12**
*US Eng. / M / Child	62.31	51.26	**51.99**
US Eng. / M / Adult	68.58	52.63	**55.46**
French / F / Senior	72.53	54.82	**56.35**
US Spanish / F / Adult	73.71	48.07	**48.37**

Y. Jiao et al., "Universal Neural Vocoding with Parallel WaveNet", ICASSP 2021

Universal Neural Vocoding

Comparison with speaker-dependent vocoders

MUSHRA results per style

MUSHRA	Recording	SDPW	UPW
All Internal	69.68	57.92	**58.70**
Emotional	71.59	60.74	**61.40**
Neutral	69.13	58.53	**58.73**
Conversational	58.65	43.54	**47.61**
Long-form reading	68.60	**56.69**	55.46
News briefing	75.24	56.29	**59.86**
Singing	71.94	49.96	**56.87**

Y. Jiao et al., "Universal Neural Vocoding with Parallel WaveNet", ICASSP 2021

Data Reduction

1. We augment data via voice conversion (VC) by leveraging recordings in the desired speaking style from other speakers.
2. We use that synthetic data on top of the available recordings to train a TTS model.
3. We fine-tune (FT) that model on the target recordings.

Recordings
Non-data reduced scenario (full target data; 45 min, 1.5h or 5h)
Our methodology, using limited target data (DR; 30 min), voice-converted data (VC) and fine-tuning (FT)

		Naturalness	Speaker similarity
SMALL - Reducing 45 min to 30 min			
4 speakers	Recs	78.80±0.73	68.10±0.65
	non-DR	51.78±0.76	65.61±0.67
	DR+VC+FT	**56.86**±0.72	**66.80**±0.66
MEDIUM - Reducing 1.5h to 30 min			
2 speakers	Recs	79.86±1.02	69.78±0.89
	non-DR	56.58±1.04	**67.45**±0.91
	DR+VC+FT	**58.72**±1.02	67.41±0.90
LARGE - Reducing 5h to 30 min			
2 speakers	Recs	81.65±1.00	67.83±0.94
	non-DR	55.16±1.06	66.25±0.95
	DR+VC+FT	**59.39**±0.97	**66.72**±0.93

G. Huybrechts et al., "Low-resource expressive text-to-speech using data augmentation", ICASSP 2021

Data Reduction

Newscaster

Baseline vs Our Data Reduction methodology

30 min of target recordings

30 min of target recordings

Conversational

Baseline vs Our Data Reduction methodology

5h of target recordings

30 min of target recordings

1.5h of target recordings

30 min of target recordings

G. Huybrechts et al., "Low-resource expressive text-to-speech using data augmentation", ICASSP 2021

Alexa NTTS

Natural

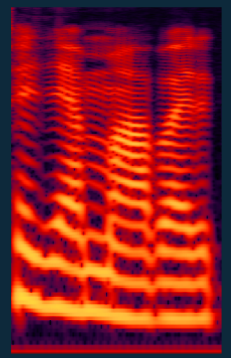
Expressive

Neural Text-to-Speech

PHONEMES &
LINGUISTIC FEATURES

ACOUSTIC MODELING

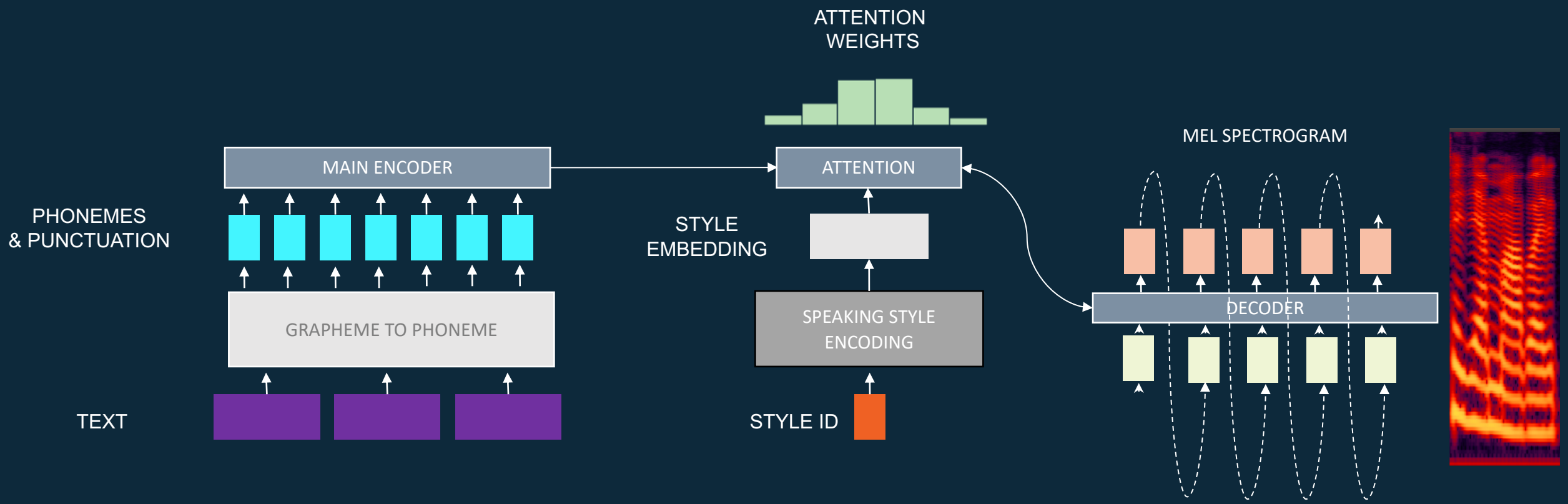
MEL
SPECTROGRAMS



(Sequence-to-sequence network)

- Focused on speech representation
  - Speaker and style specific

Speaking style



Newscaster TTS

Introducing Amazon's first custom electric delivery vehicle

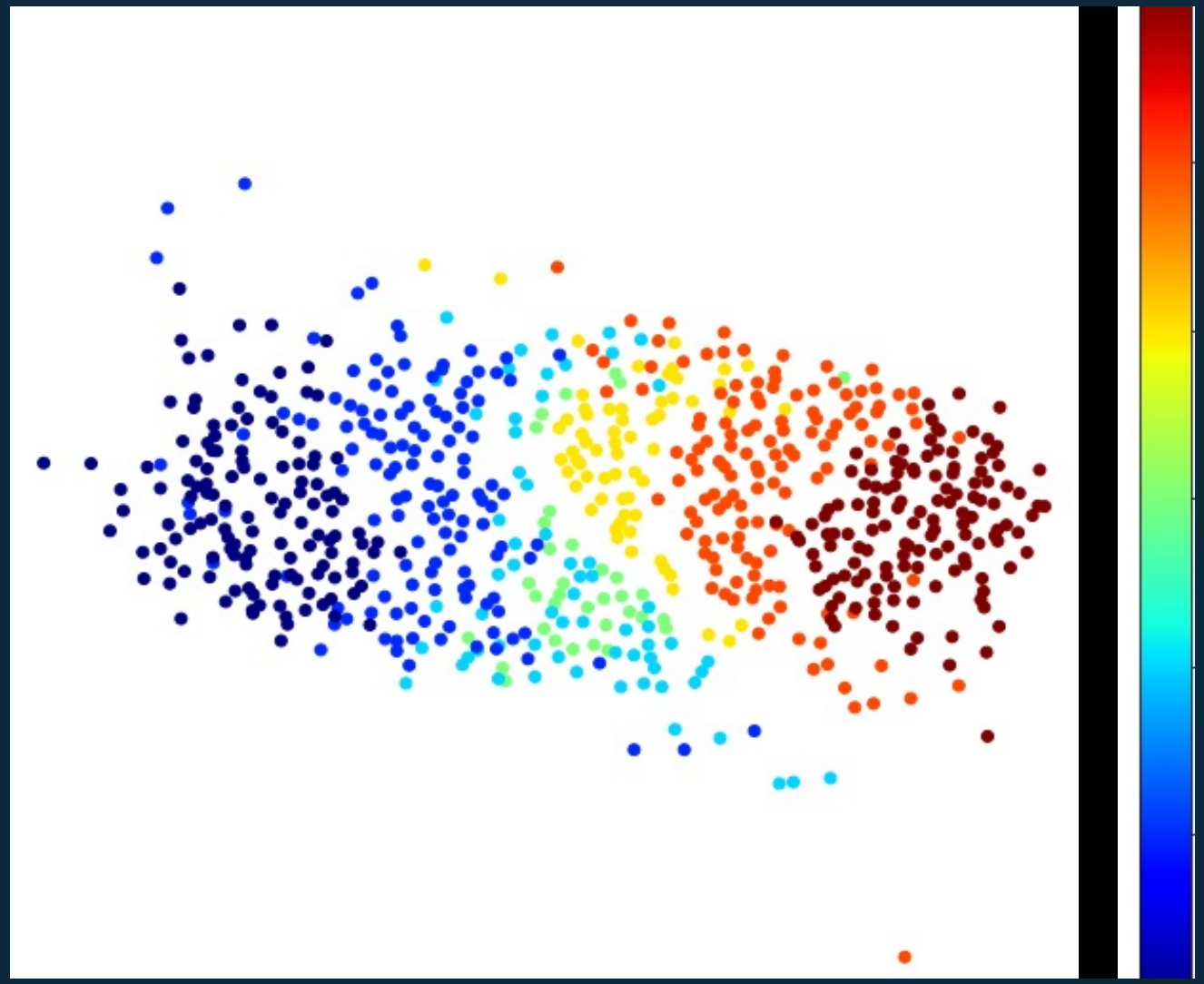
"Amazon has revealed its first custom electric delivery vehicle, designed and built in partnership with Rivian, and expects to have 10,000 of the new vans on the road delivering to customers as early as 2022."

Neutral

Newscaster

N. Prateek et al., "In Other News: A Bi-style Text-to-speech Model for Synthesizing Newscaster Voice with Limited Data", NAACL 2019

Emotional Speech



Excited, High

Excited, Medium

Excited, Low

Neutral

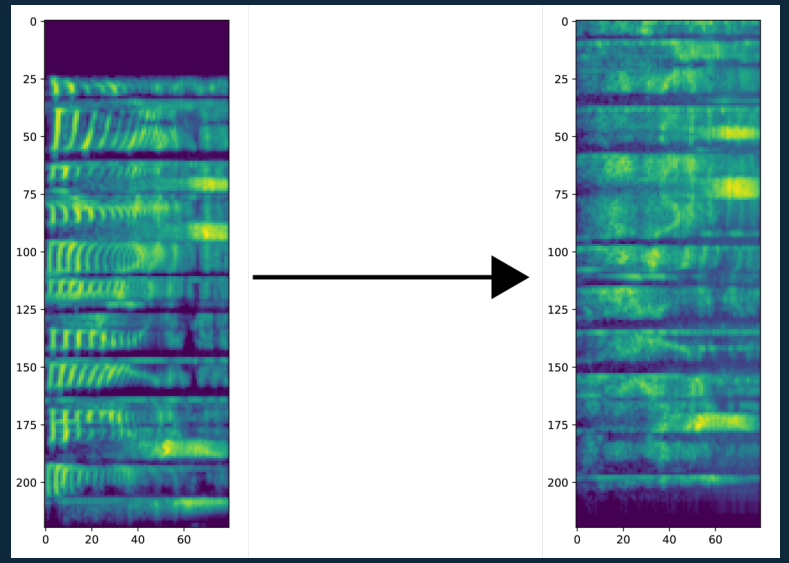
Disappointed, Low

Disappointed, Medium

Disappointed, High

Whisper Mode

Idea:
Transform a spectrogram of normal voice into a spectrogram of whispered voice through Machine Learning

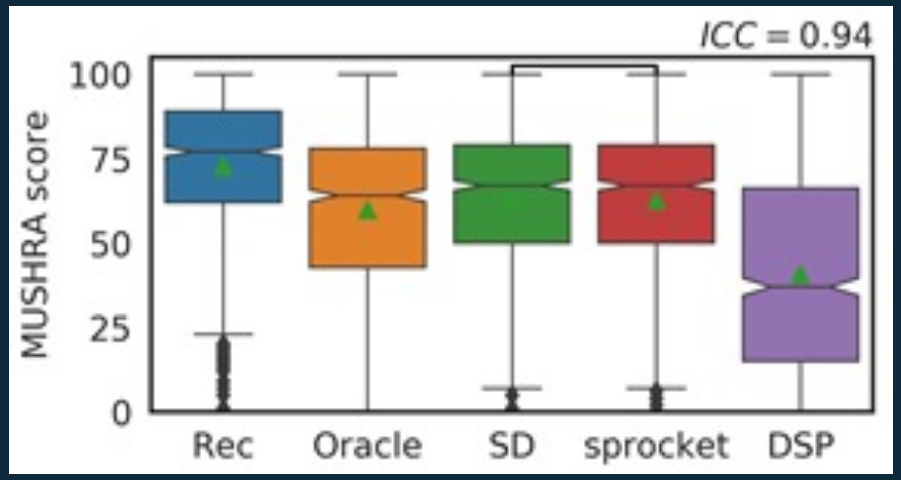


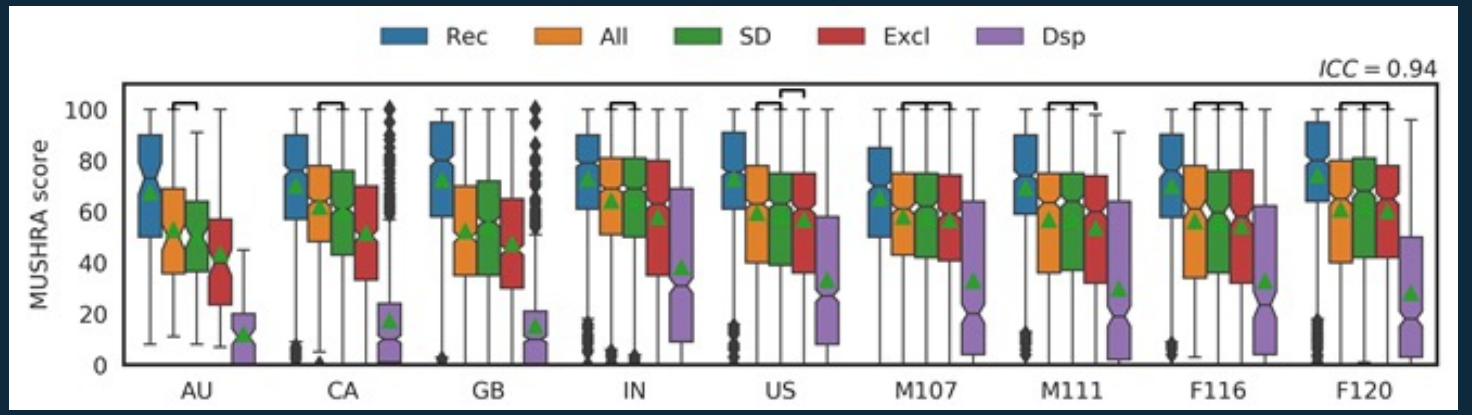
M. Cotescu et al., "Voice Conversion for Whispered Speech Synthesis", Sig. Proc. Letters 2019

Whisper Mode

Based on DSP

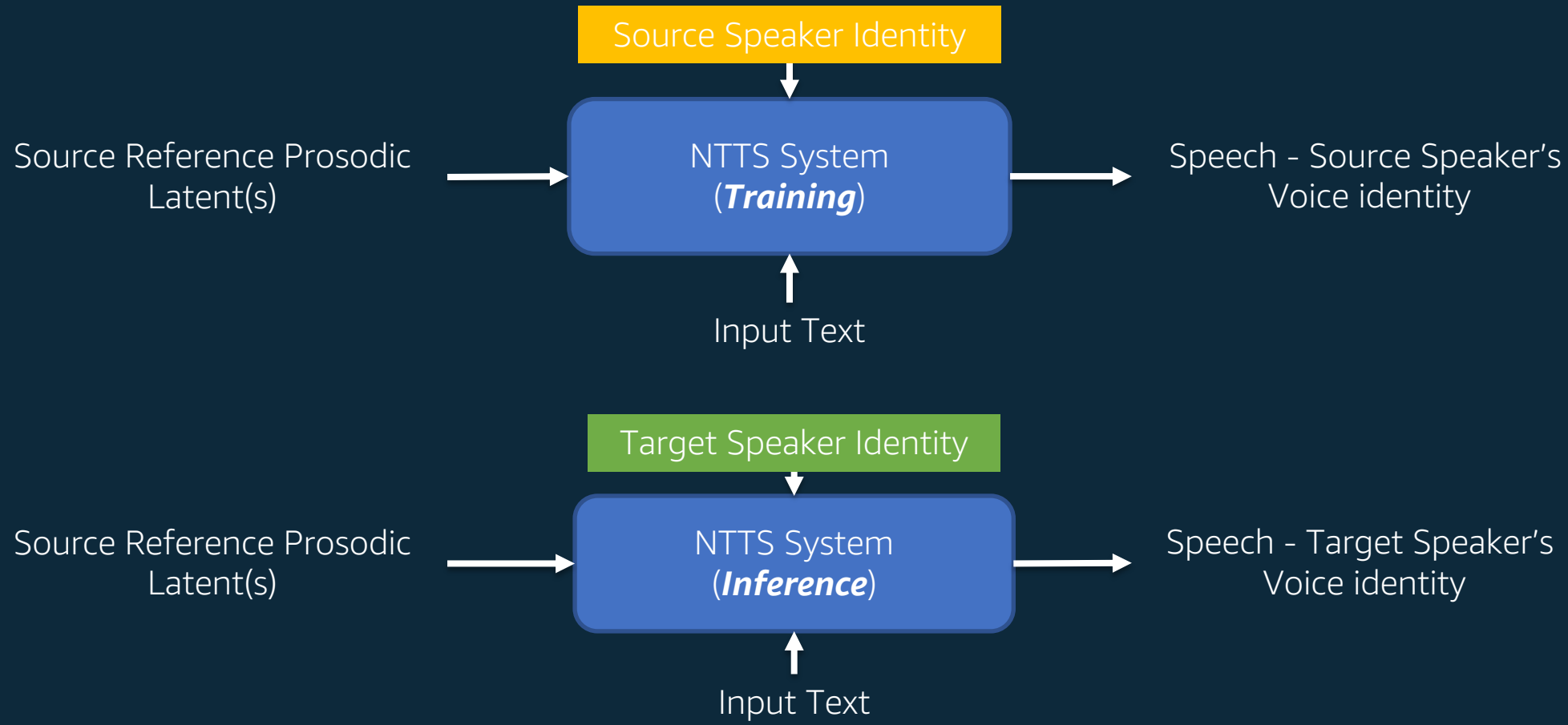
Based on Machine Learning





M. Cotescu et al., "Voice Conversion for Whispered Speech Synthesis", Sig. Proc. Letters 2019

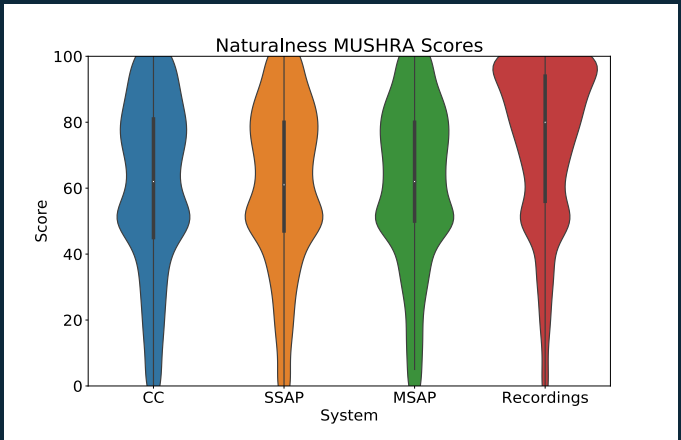
Prosody Transfer

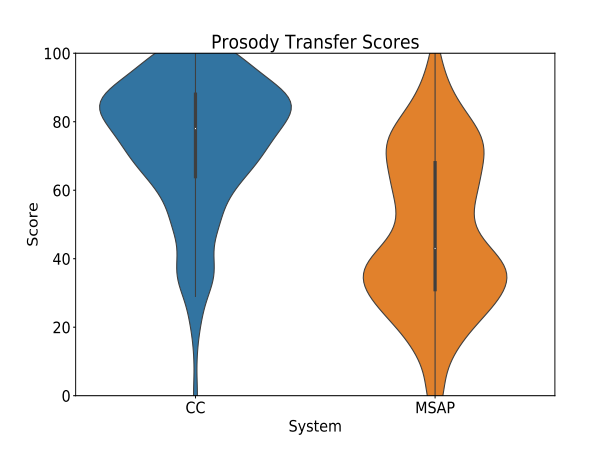


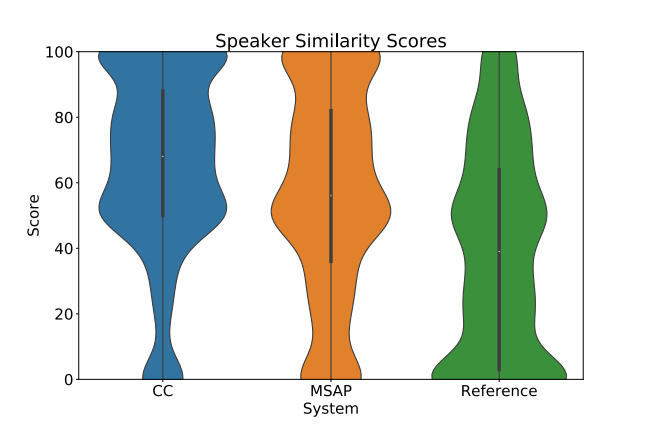
S. Karlapati et al., "CopyCat: Many-to-Many Fine-Grained Prosody Transfer for Neural Text-to-Speech", Interspeech 2020

Prosody Transfer

Source Prosodic Reference	Target Speaker Identity	Synthesised Speech







S. Karlapati et al., "CopyCat: Many-to-Many Fine-Grained Prosody Transfer for Neural Text-to-Speech", Interspeech 2020

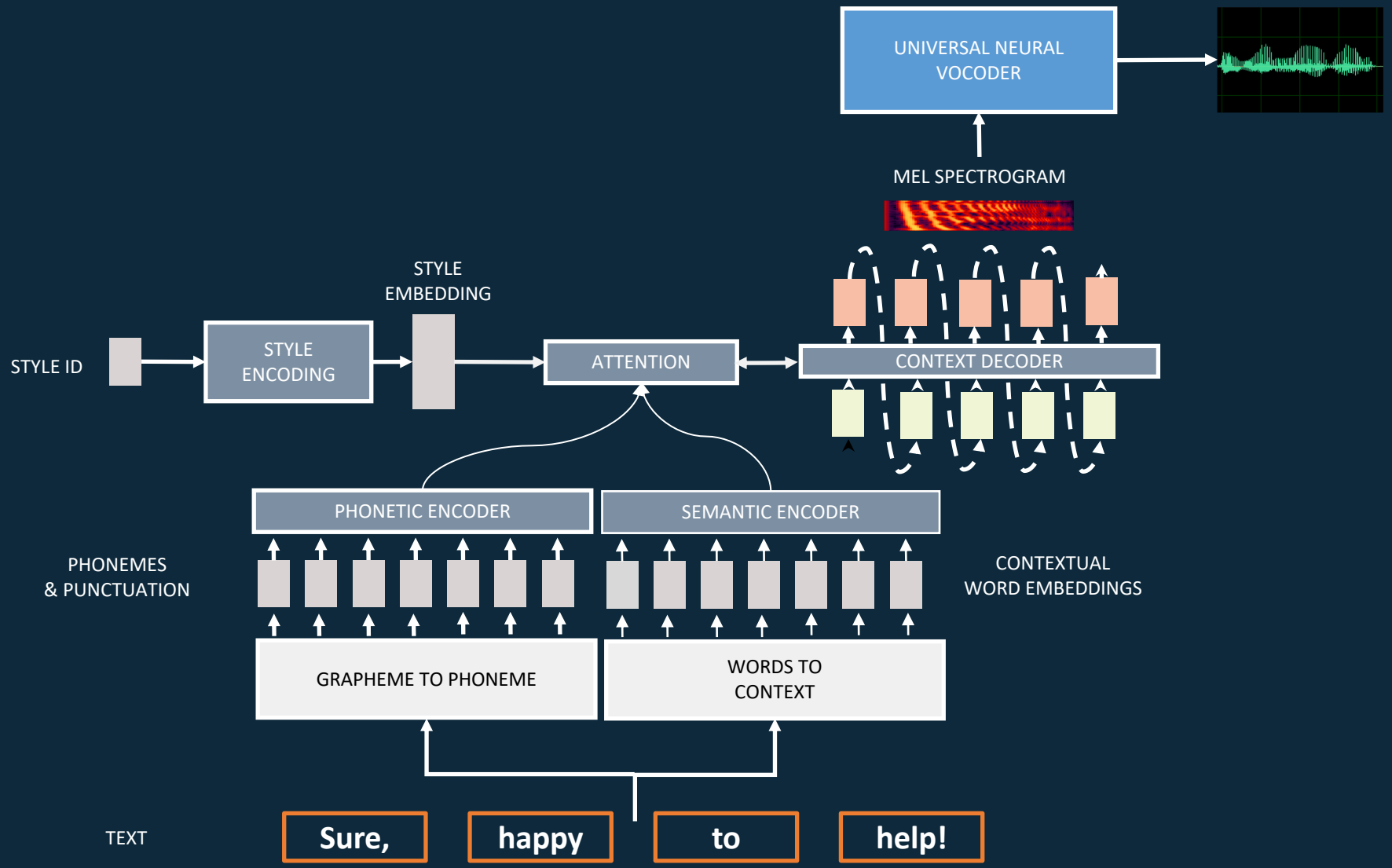
Alexa NTTS

Natural

Expressive

Aware

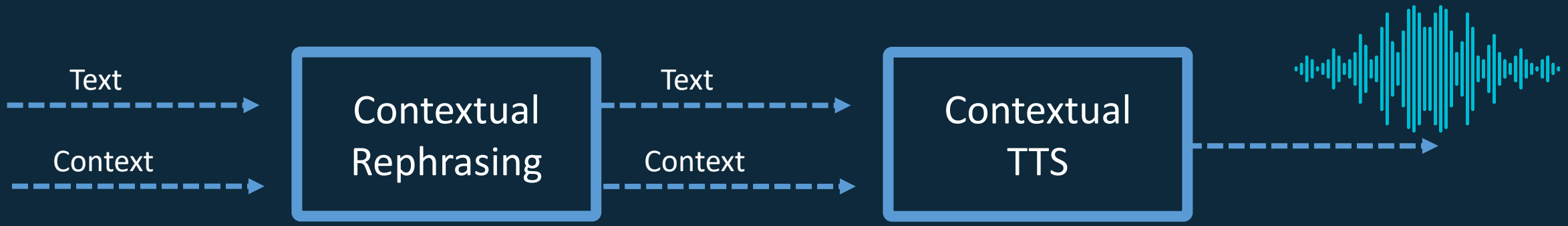
Syntactic and semantical info



S. Karlapati et al., "Prosodic Representation Learning and Contextual Sampling for Neural Text-to-Speech", ICASSP 2021

Dialogues

Input: Text + additional context information



⇒ Contextual response & diversity

Content and Context Aware