AWS re:Invent

ALX308

Making Alexa sound more human — Amazon's text-to-speech technology

Andrew Breen
Sr. Research Manager
Amazon Alexa
Amazon.com

Nikhil Sharma
Sr. Product Manager
Amazon Alexa
Amazon.com

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

- Voice strategy
- Technology behind Alexa’s voice
- Q&A

Related breakouts

- ALX201 - How developers can build natural, extensible voice conversations

- ALX309 - Alexa’s different speaking styles for developers

- ALX202 - Alexa, what can I do now?

Amazon text-to-speech (TTS)

Echo devices







Other Alexa devices







AWS AI services



Amazon Lex

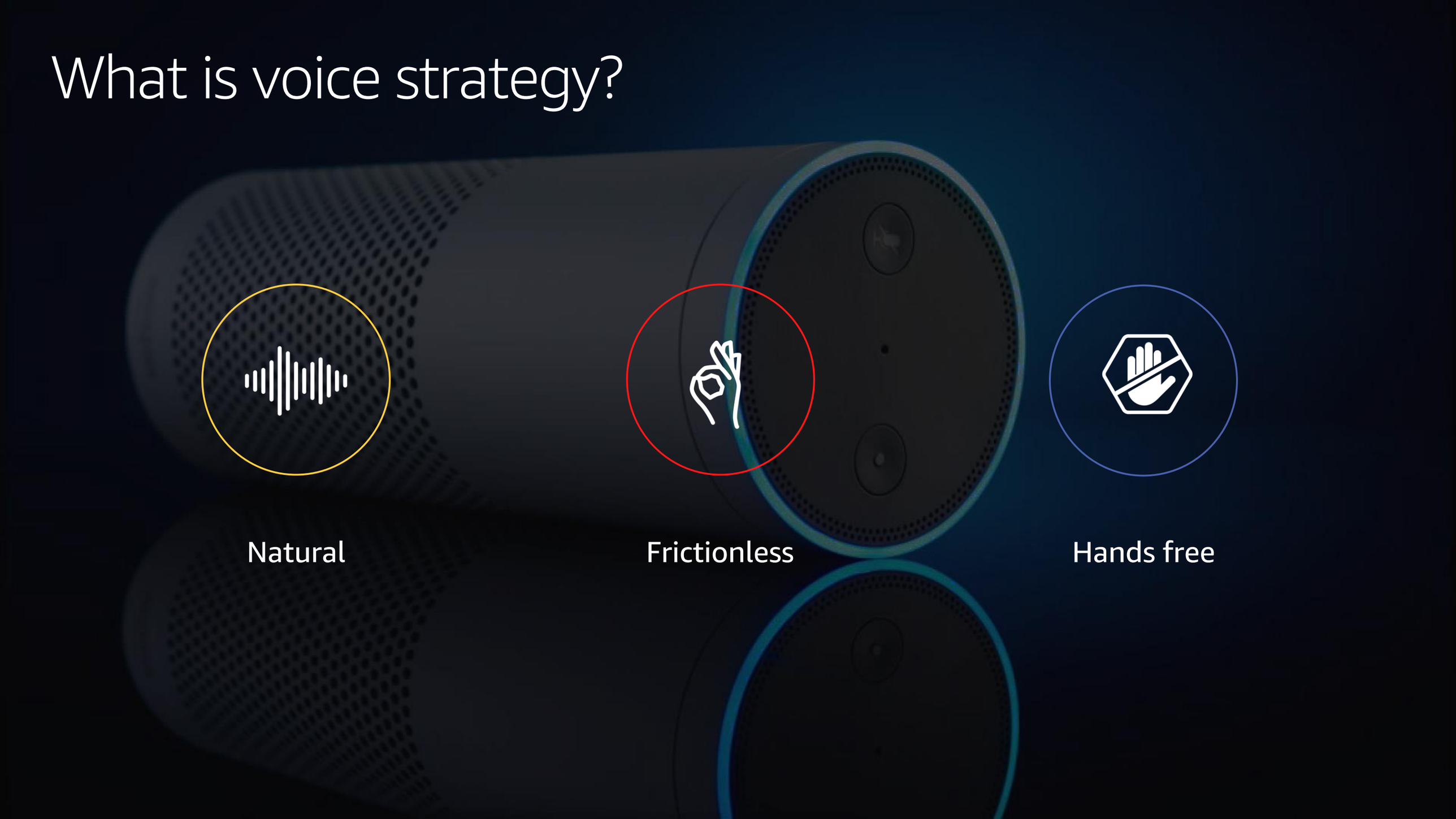


Amazon Polly

Voice strategy

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is voice strategy?

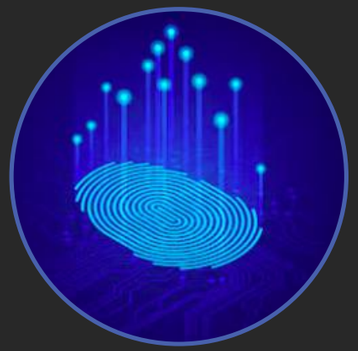


Natural

Frictionless

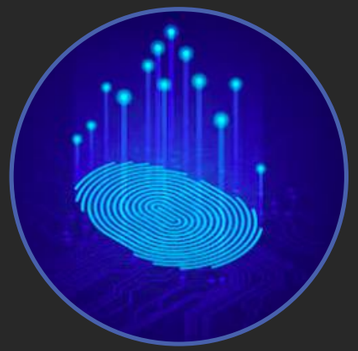
Hands free

Components of Alexa’s voice experience



Human-like

Components of Alexa's voice experience

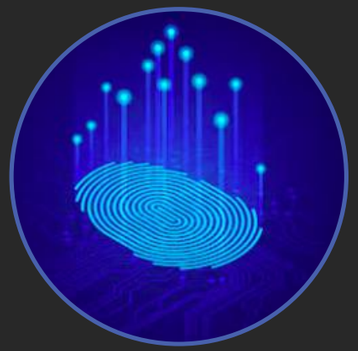


Human-like



Focus on utility

Components of Alexa's voice experience



Human-like

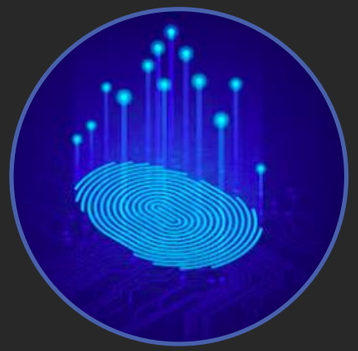


Focus on utility



Alexa has a personality

Components of Alexa's voice experience



Human-like



Focus on utility



Alexa has a personality



Always learning, always improving

Alexa's voice projection: Trusted and Secure



Alexa's voice projection: Conversational



Alexa’s voice projection: Intelligent

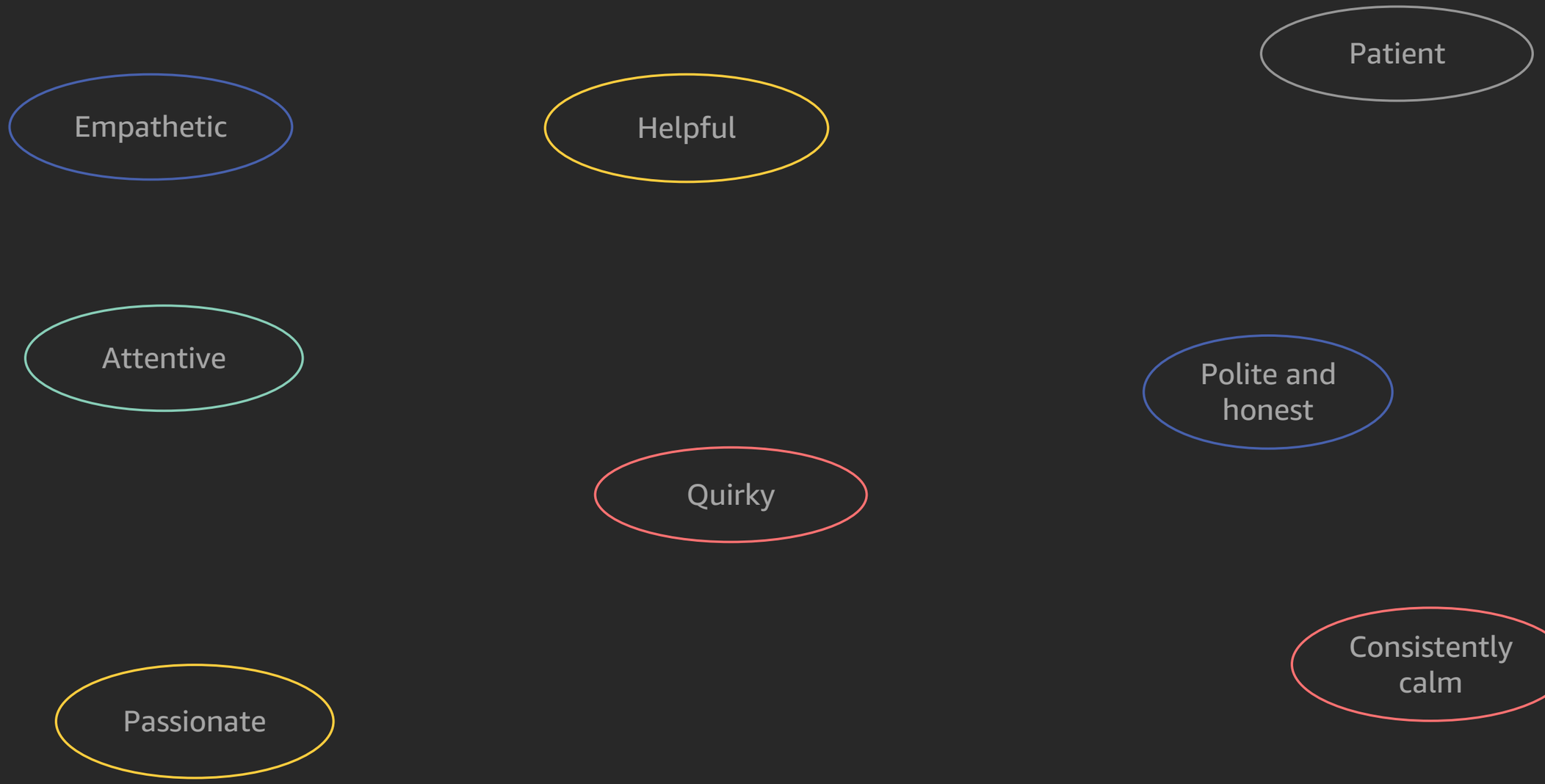


Alexa’s voice projection: Makes customers smile





Factors influencing voice selection: Personality



Factors influencing voice selection: Brand projection





Factors influencing voice selection: Gender



Factors influencing voice selection: Culture and region





Voice strategy from partners: PolicyBazaar

"Being an early adopter of technology will ultimately contribute to a better bottom line for our business. Although there are other TTS innovations, the design of Amazon Polly facilitates seamless integration into our systems. Further, its capability to understand the Indian accent also makes it an ideal choice for us."

Ashish Gupta

PolicyBazaar

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Voice strategy from partners: Inhealthcare

"In a world of innovative apps and clever technology, people sometimes forget there are many individuals who are excluded from the digital world for social, economic or health reasons. As a trusted partner of the NHS, we rely on Amazon Polly to do all the synthesized speech for our automated telephone calls. Individuals can answer questions and provide information using the telephone's keypad. Amazon Polly helps us make sure these people are being cared for."

Mike Wray

Senior Software Architect

inhealthcare

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

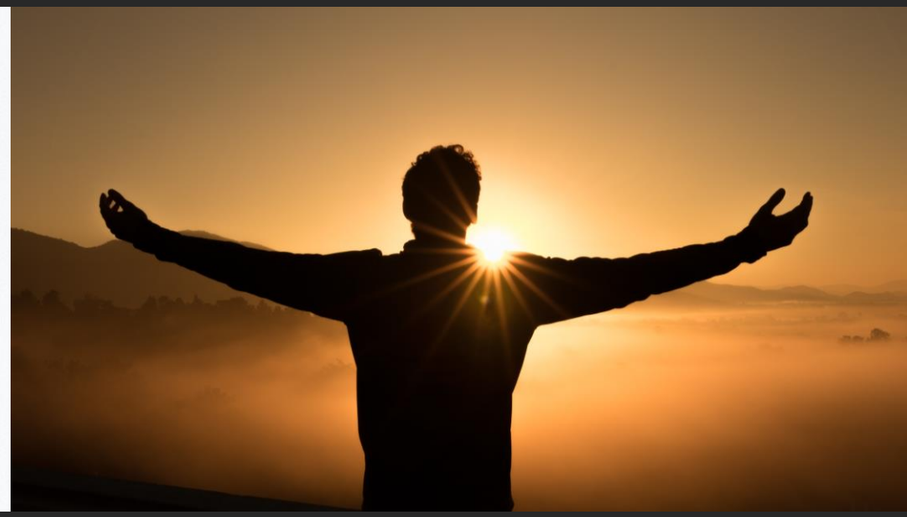
Voice strategy: Three key takeaways



Voice is important



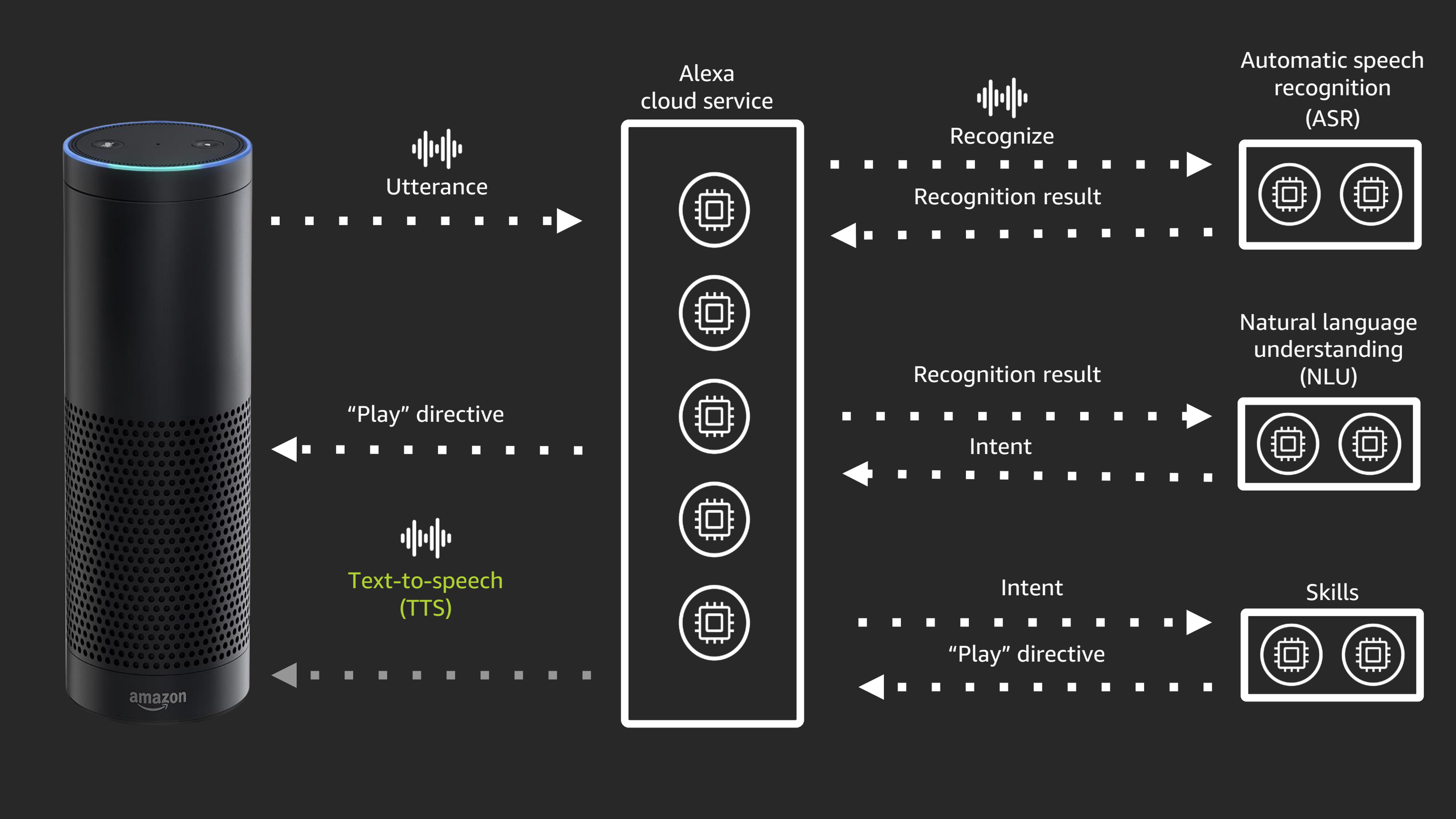
Convenience is key



Lifelike experiences

The technology behind Alexa's voice

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Alexa
cloud service
Utterance
“Play” directive
Text-to-speech
(TTS)
Recognize
Recognition result
Automatic speech
recognition
(ASR)
Recognition result
Intent
Natural language
understanding
(NLU)
Intent
“Play” directive
Skills
amazon

Brief history of 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



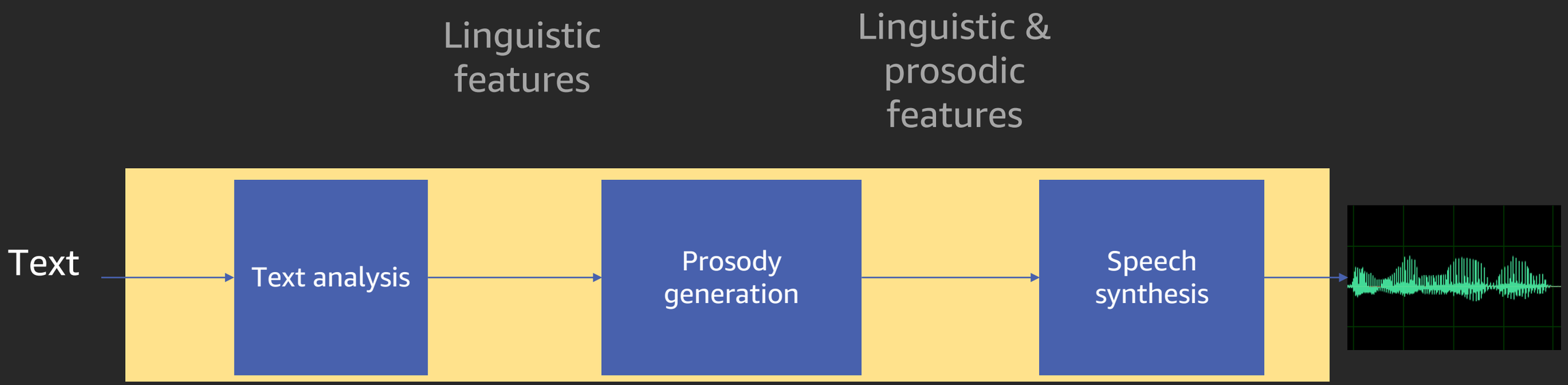
2018

NEURAL TTS

First applications using Alexa neural TTS

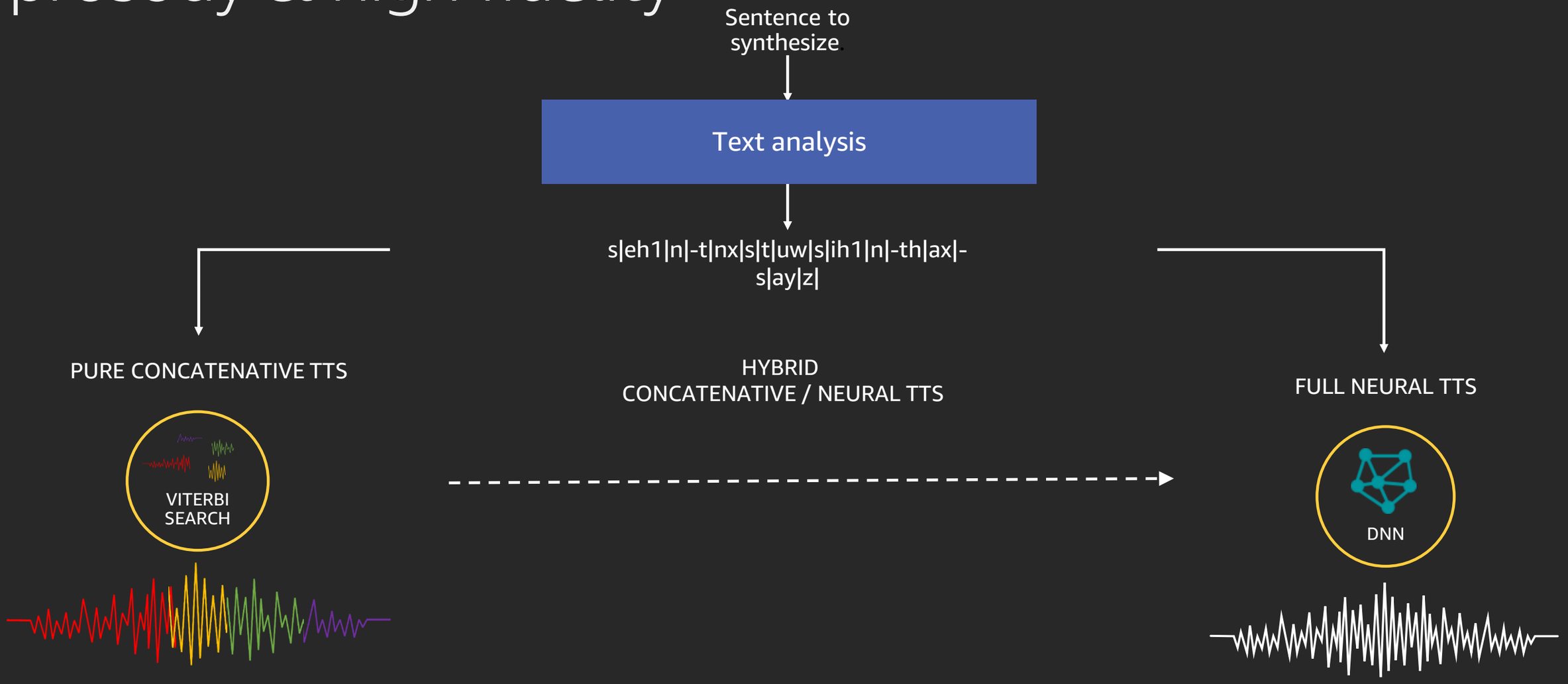


Components of a TTS system

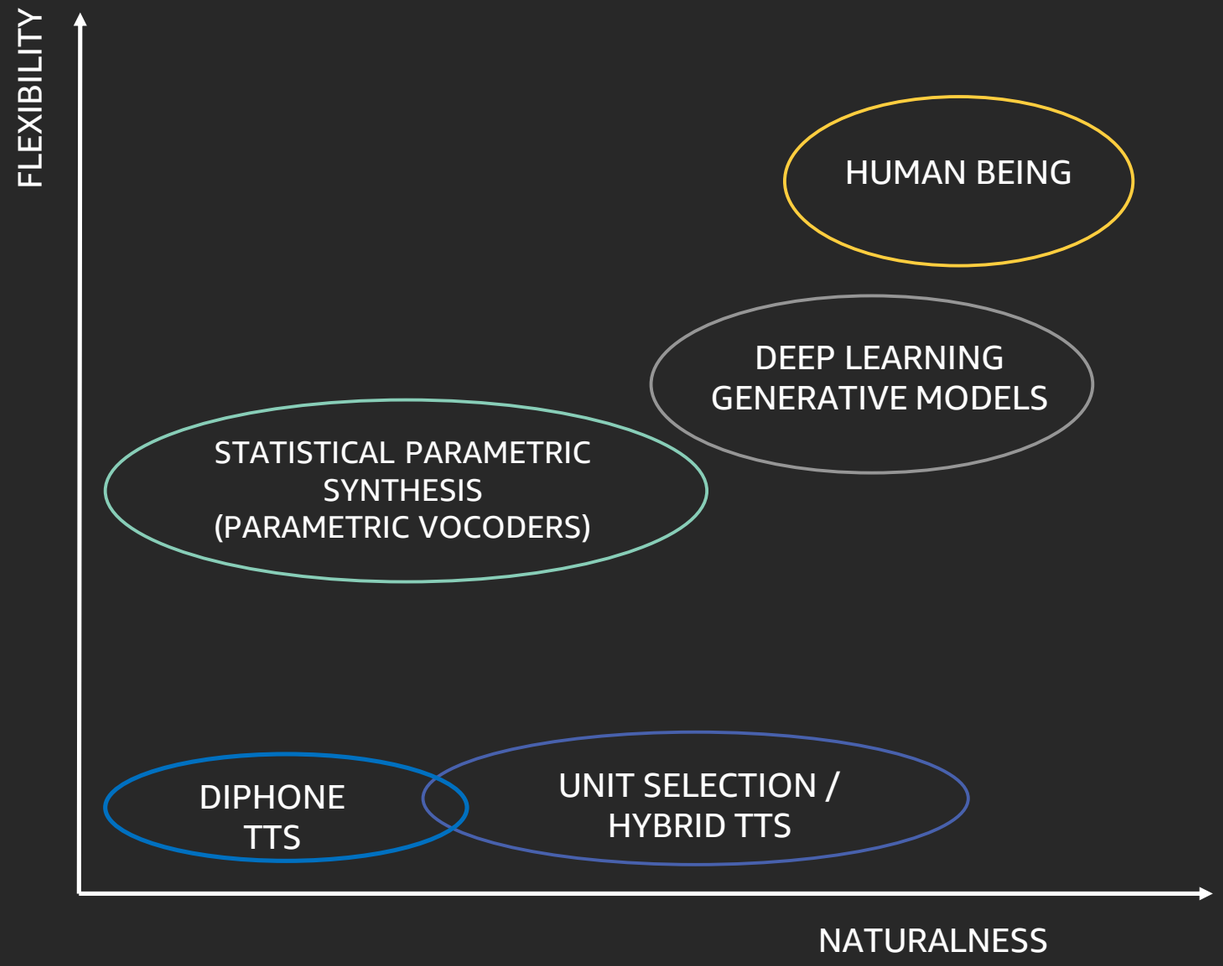


- Extract linguistic knowledge
- Syntax, part of speech, sounds of language
- Mathematical models of style
- Speech rate, pitch, & phrasing
- Generation of acoustic signal

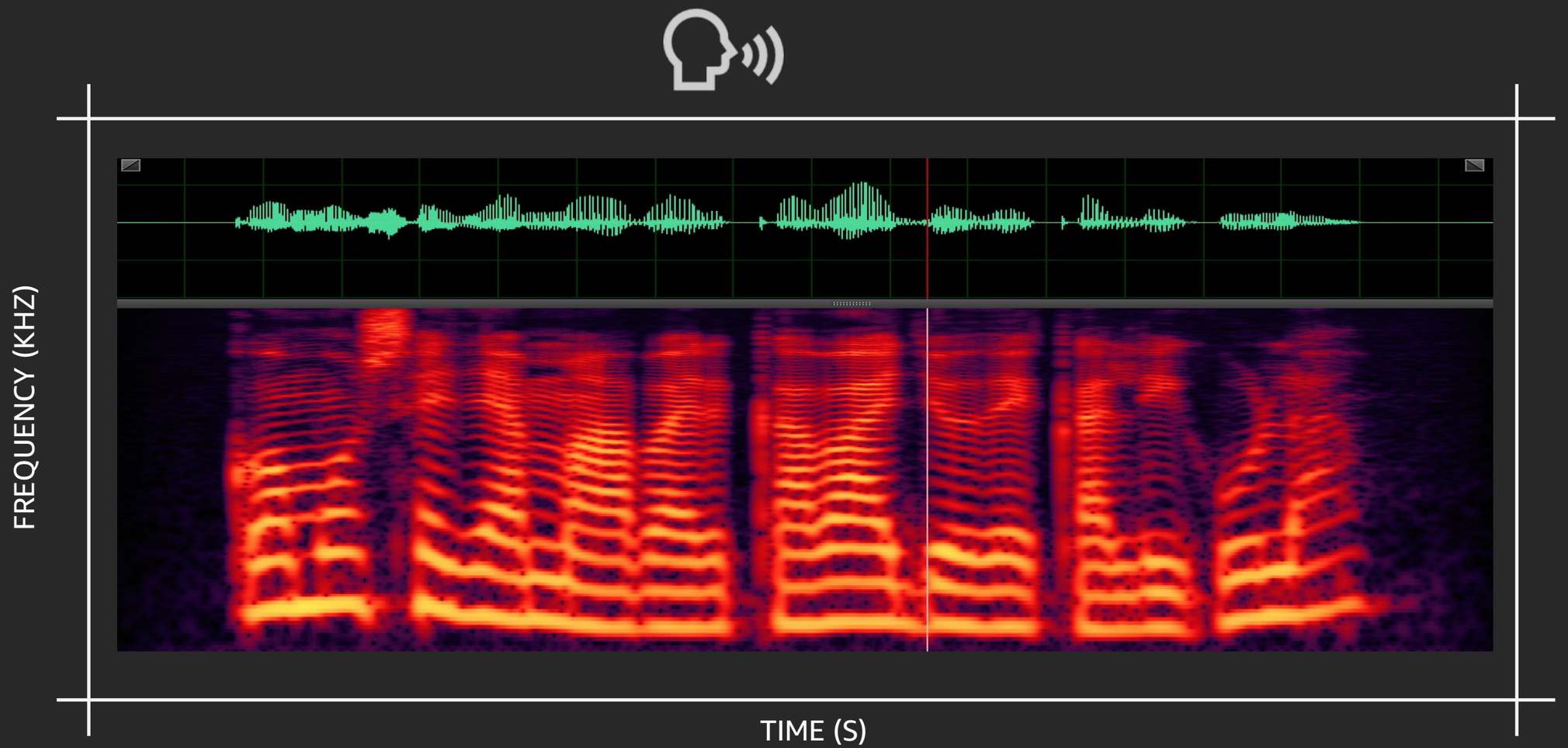
Neural text-to-speech (NTTS) brings together natural prosody & high fidelity



Why NTTS?


FLEXIBILITY
HUMAN BEING
DEEP LEARNING
GENERATIVE MODELS
STATISTICAL PARAMETRIC
SYNTHESIS
(PARAMETRIC VOCODERS)
DIPHONE
TTS
UNIT SELECTION /
HYBRID TTS
NATURALNESS

How speech is represented



How does it work?

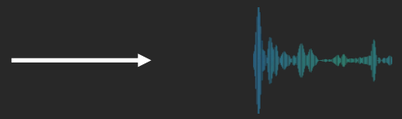
PHONEMES &
FEATURES

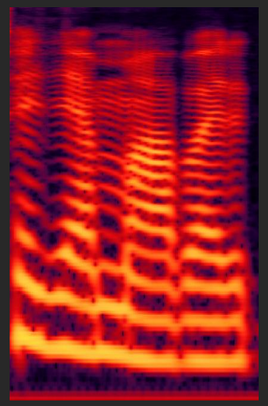
CONTEXT
GENERATION

(Sequence-to-sequence network)

- Focused on speech representation
  - Speaker and style specific

MEL
SPECTROGRAMS





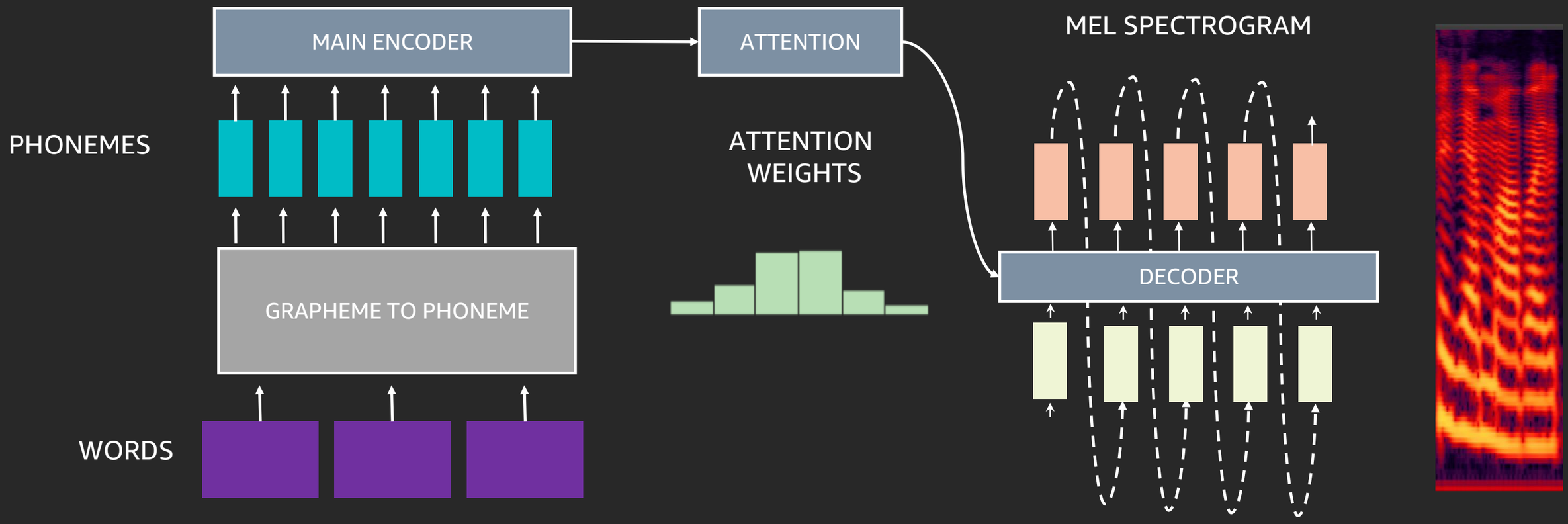
SPEECH

SPEECH PRODUCTION

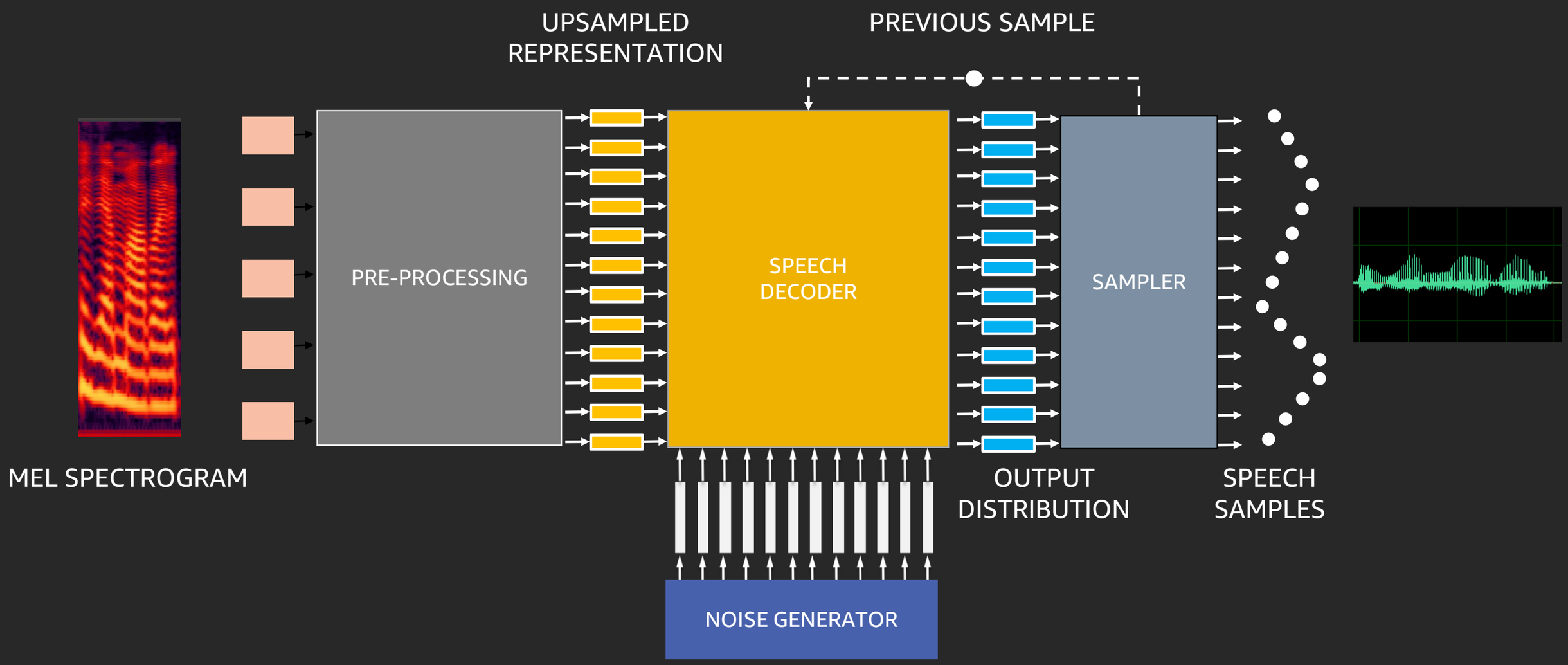
(Neural vocoder)

- Focused on hi-fi speech signal
- Independent of speaker or style

Deep dive: Context generation

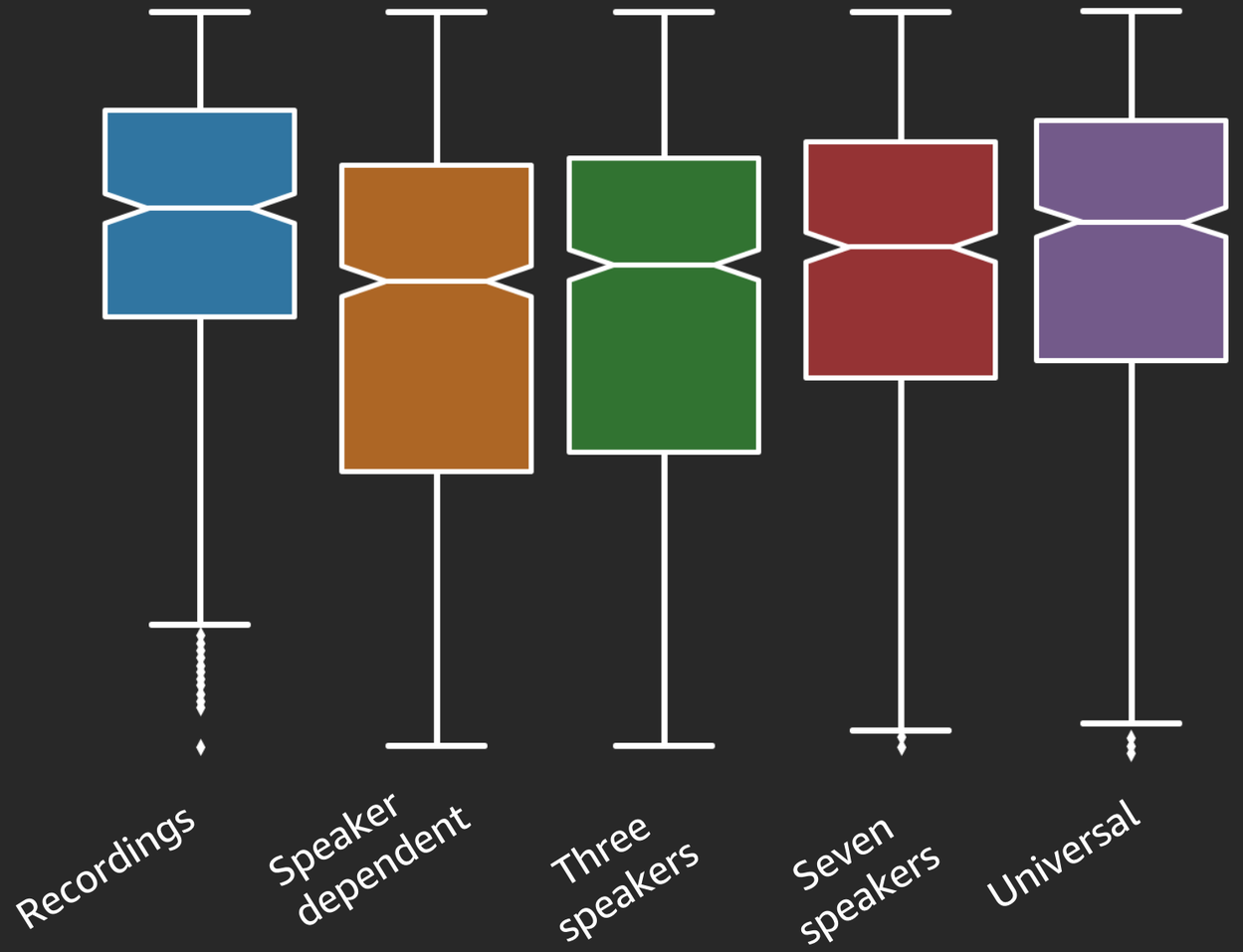


Deep dive: Neural vocoder



Naturalness





Amazon universal neural vocoder

- Hi-fi audio
- Speaker/language independent model trained on multiple voices from AWS voices portfolio (>200 hours)







How it sounds in Alexa

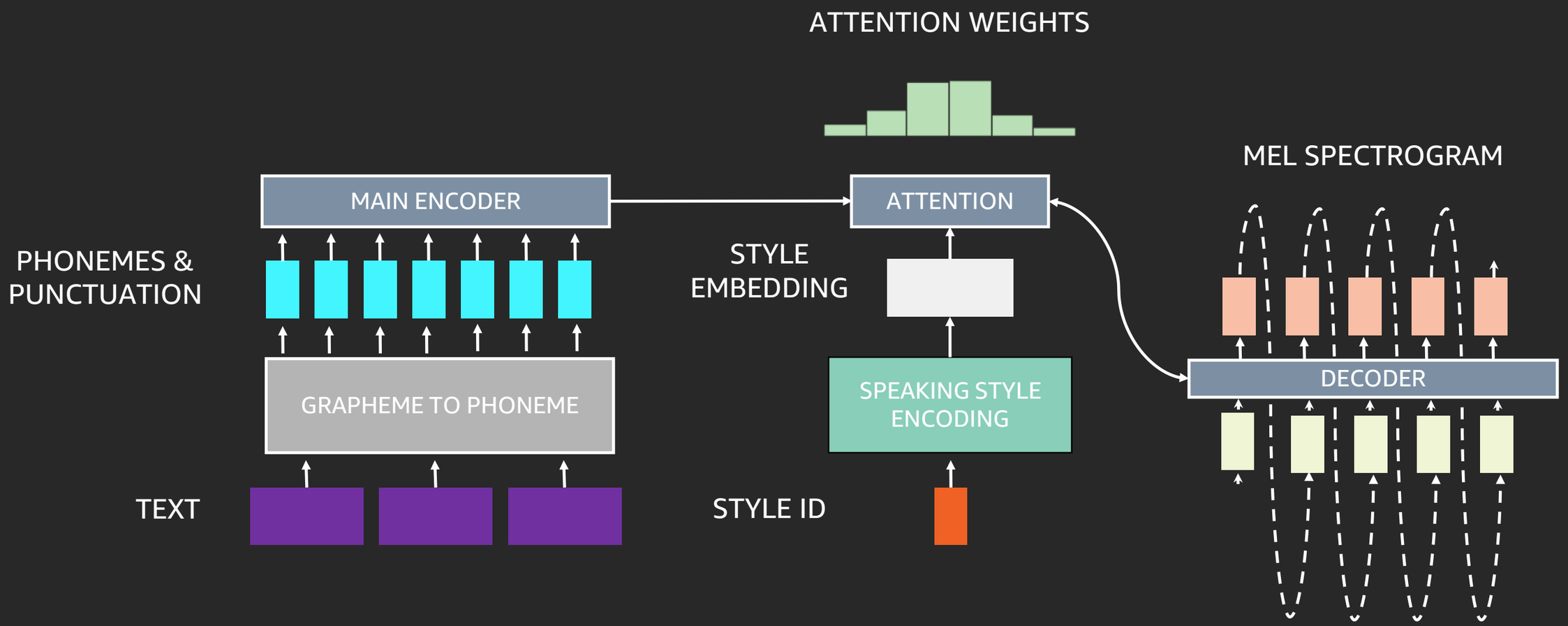
WIKIPEDIA: PIKE PLACE MARKET

“Pike Place Market is a public market overlooking the Elliot Bay waterfront in Seattle. The Market opened August 17, 1907, and is one of the oldest continuously operated public farmers’ markets in the United States. It is a place of business for many small farmers, craftspeople, and merchants.”

CONCATENATIVE

NTTS

Building a speaking style



Improving experiences through NTTS

SAN FRANCISCO'S $2.2 BILLION 'GRAND CENTRAL' TERMINAL OPEN

"Spanning three city blocks, San Francisco's $2 billion transit terminal is open after nearly a decade under construction. Coined the "Grand Central of the West," the new Salesforce Transit Center near the heart of downtown is expected to accommodate 100,000 passengers each weekday, and up to 45 million people a year. It welcomed its first passengers and visitors this month."

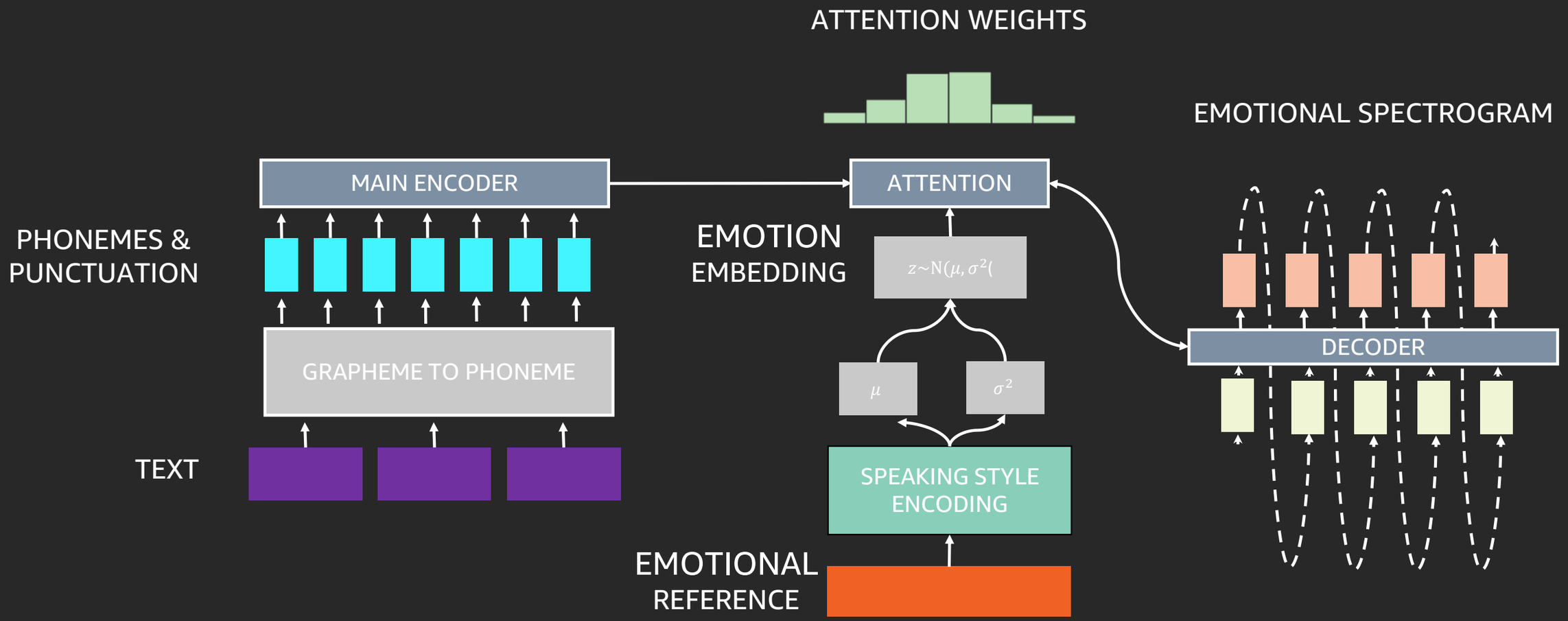


CONCATENATIVE

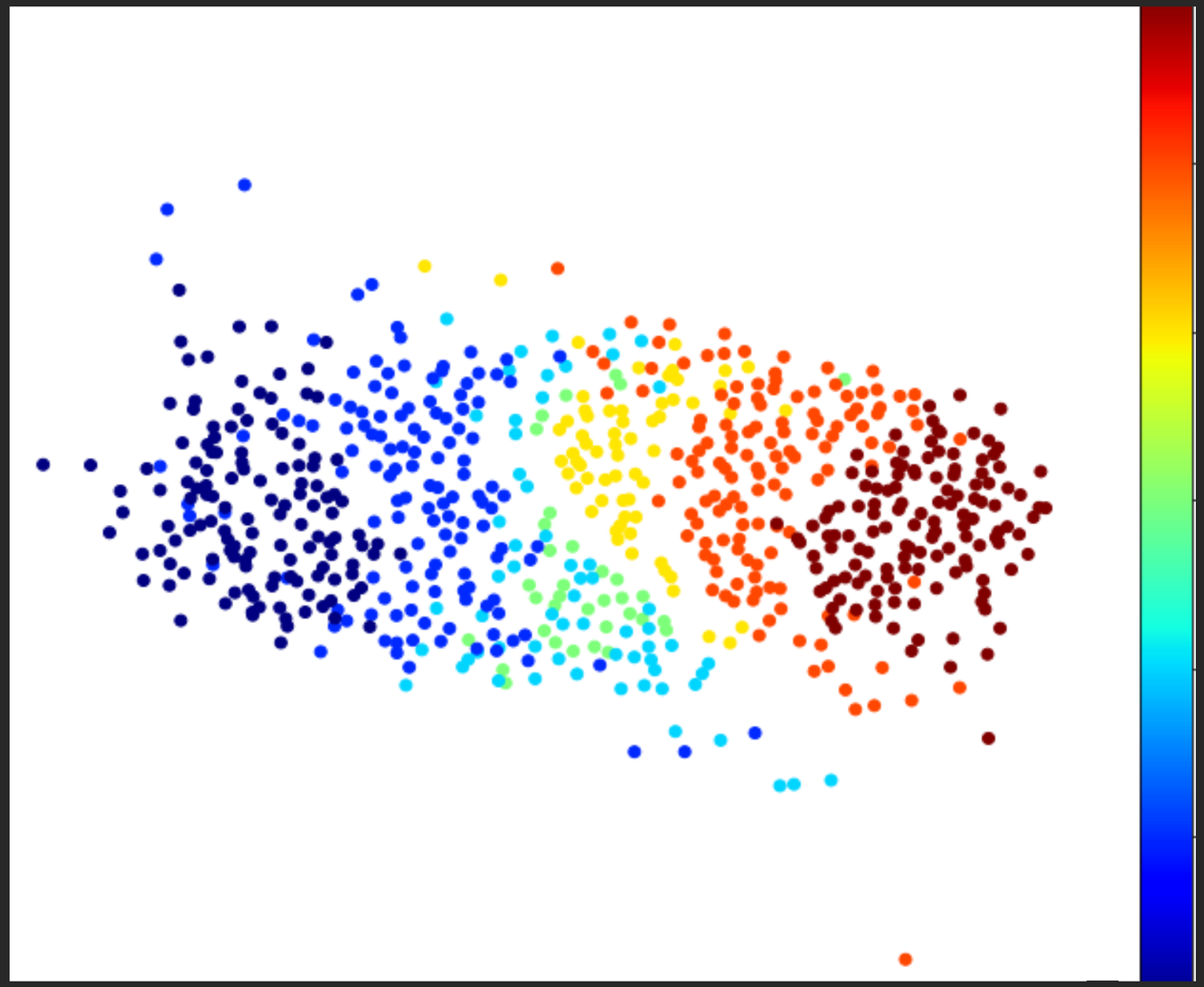


NEWSCASTER NTTS

Incorporating emotions to NTTS



Variations of emotions in Alexa's voice



Excited, High

Excited, Medium

Excited, Low

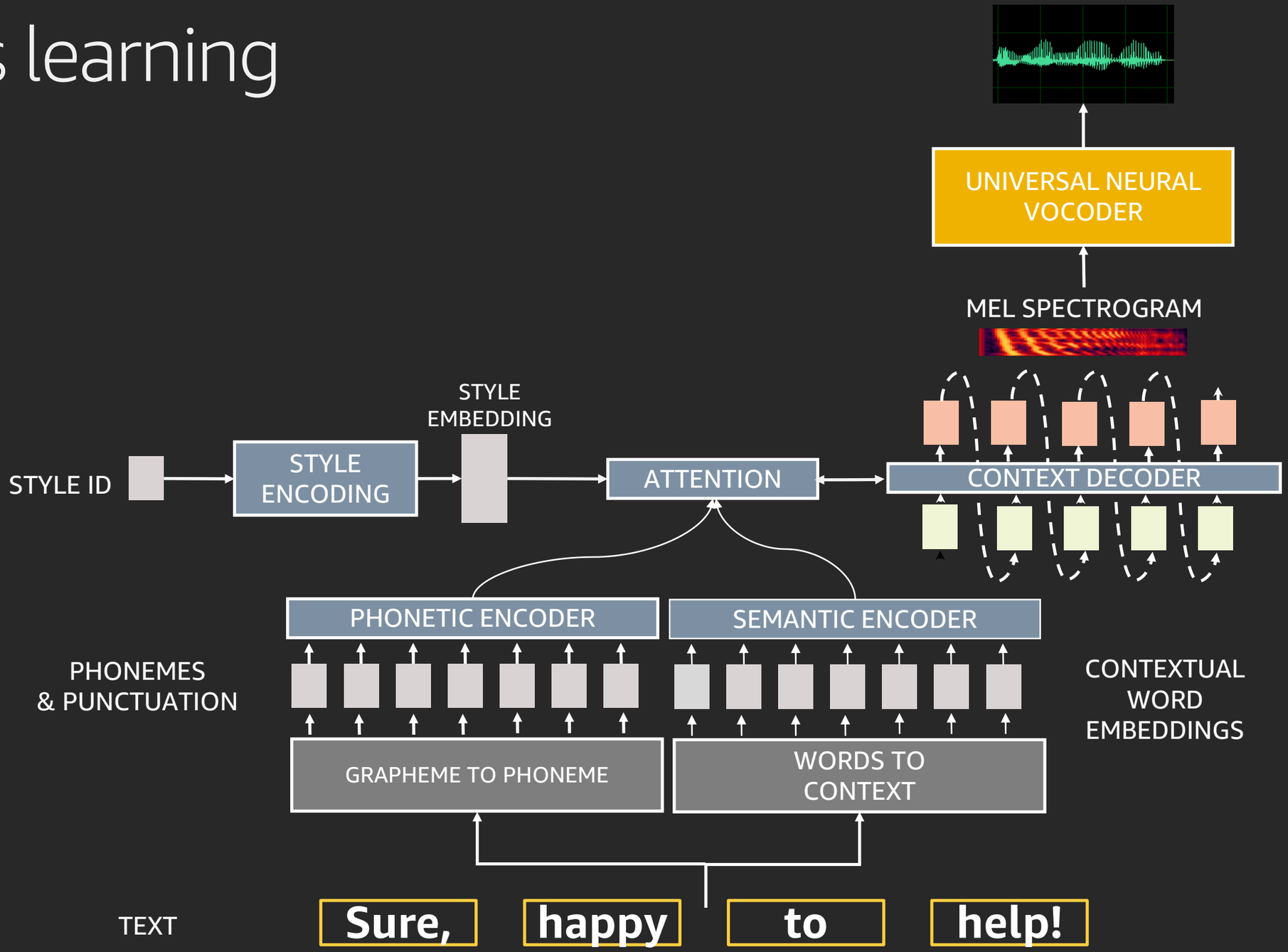
Neutral

Disappointed, Low

Disappointed, Medium

Disappointed, High

Always learning



Contextual word embeddings with NTTS

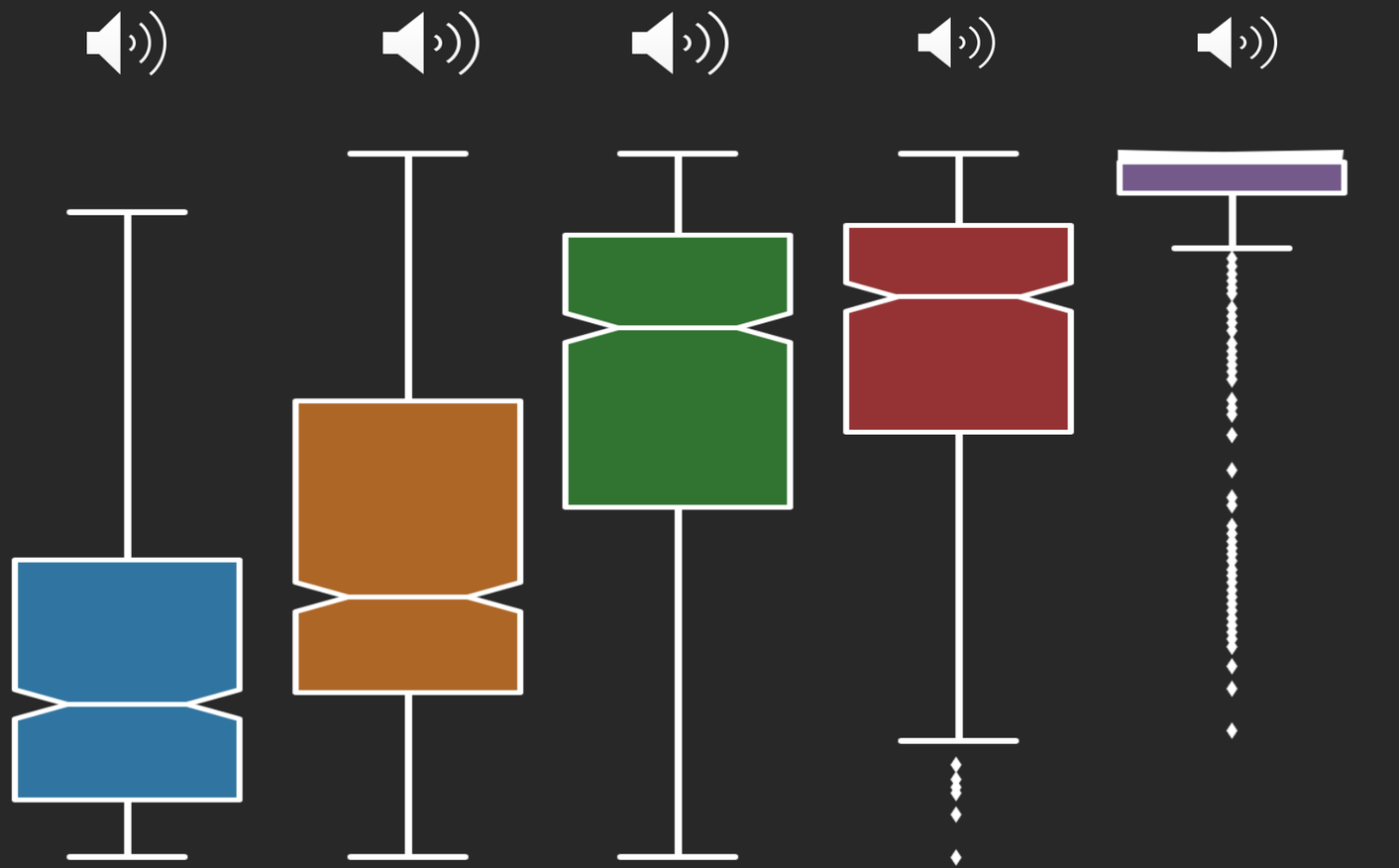
Contextual word embeddings allow us to adapt speech to match the syntax and semantics of the text

There are two kinds of such events

Without contextual word embeddings 

With contextual word embeddings 

Subjective evaluation of newscaster style



Alexa NTTS



Natural



Expressive



Aware

Learn Alexa skills with AWS Training and Certification

Build and validate skills to help you build solutions for 100+ million Alexa-enabled devices



Free digital courses cover topics related to Alexa skill-building, including Alexa for Developers and Designing for Conversation



Validate expertise with the
AWS Certified Alexa Skill Builder - Specialty exam

Visit aws.amazon.com/certification/certified-alexa-skill-builder-specialty

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws training and certification

Thank you!

AWS re:Invent

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws



Please complete the session survey in the mobile app.

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.