
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

ALX202

Advances in machine learning that make Alexa sound human-like

Nikhil Minotra (he/him)

Sr. Product Manager-Tech
Alexa Text-To-Speech
Amazon

Trevor Wood (he/him)

Applied Science Manager
Alexa Text-To-Speech
Amazon

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Introduction

- Why should voice AIs sound human-like?
- What do good human conversations sound like?

Text-to-speech (TTS) technical background

- Alexa stack
- Machine learning architecture overview

Machine learning to create more human-like speech

- Contextual speech
- Relatable speech
- Personalized speech

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Interacting with Alexa: As simple as speaking to another person



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scale of adoption of voice AI

- According to reports, ~37% of the US population, or 123.4 million people, will use voice assistants in 2022
- Voice AI devices > world population by 2024 (> 8 billion)

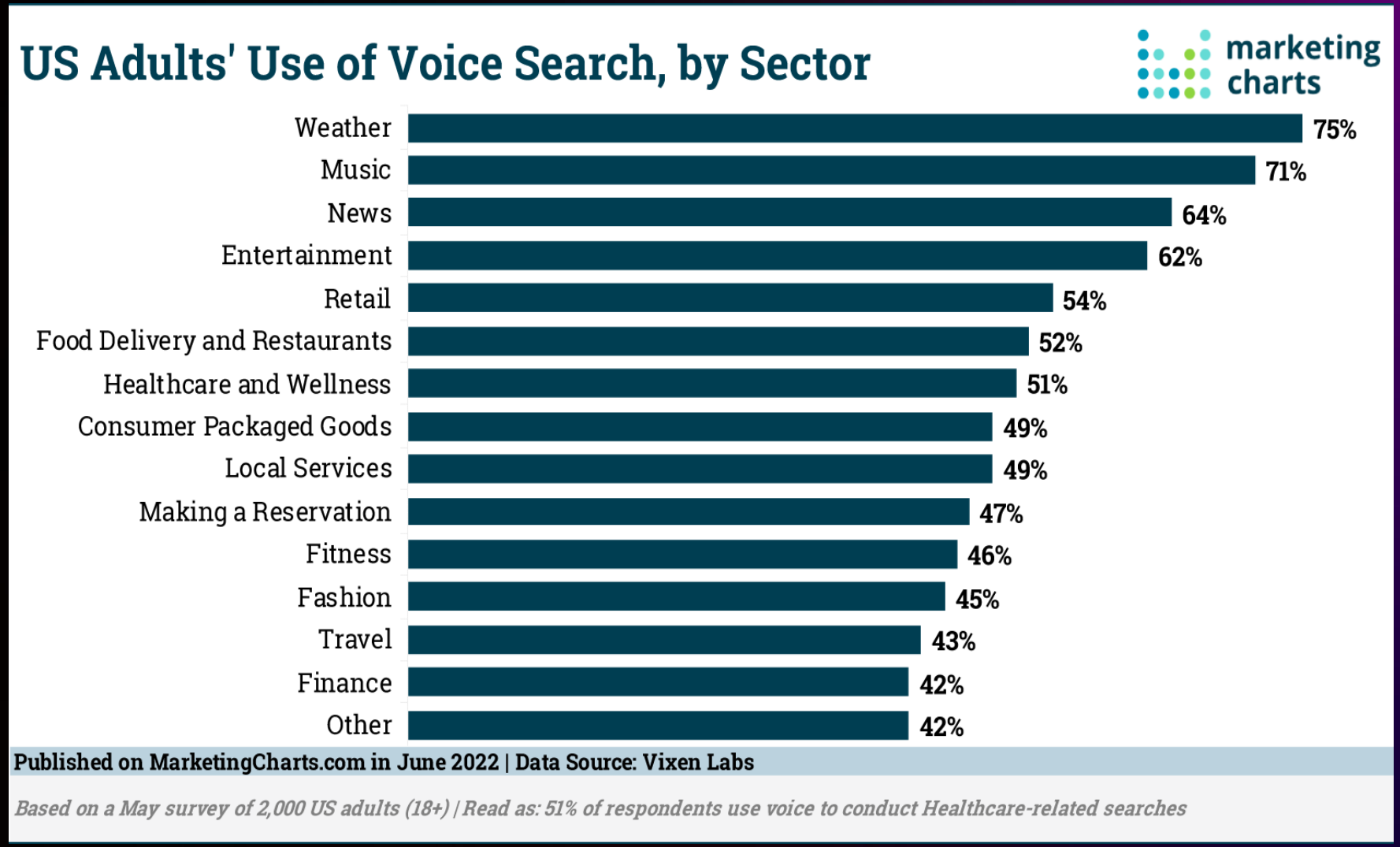


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data sources: https://www.statista.com/topics/5572/virtual-assistants/#topicHeader__wrapper

Breadth of adoption of voice AI

Used for multiple interactions, across different times and spaces of your day



Source: https://www.marketingcharts.com/digital/non-mobile-connected-devices-226268

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why should AI voices sound human-like?

- Humanized AI voices improve customer delight and usage
- As voice assistants become **more powerful**, **easier to use**, and able to **understand context** far better, more people will turn to voice search and virtual assistants for help with their everyday tasks
- As the realism and perceived human-likeness of a voice increases, its acceptance for social applications (e.g., care, companionship) should increase as well



Source: https://www.phonexia.com/blog/the-future-of-voice-assistants; Robot Voices in Our Daily Lives: Vocal Human-Likeness and Application Context as Determinants of User Acceptance. Schreibelmayr and Mara, Frontiers in Psychology, 2022.

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What do we want Alexa to be?

- Our vision is for Alexa to become an **indispensable** assistant, a **trusted** advisor, and a **fun** and **caring** companion for our customers

- Realizing this vision requires Alexa to bring the best of machine-intelligence capabilities together with the best of human-intelligence capabilities – such as perception, learning, reasoning, and interaction

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What do good human conversations sound like?

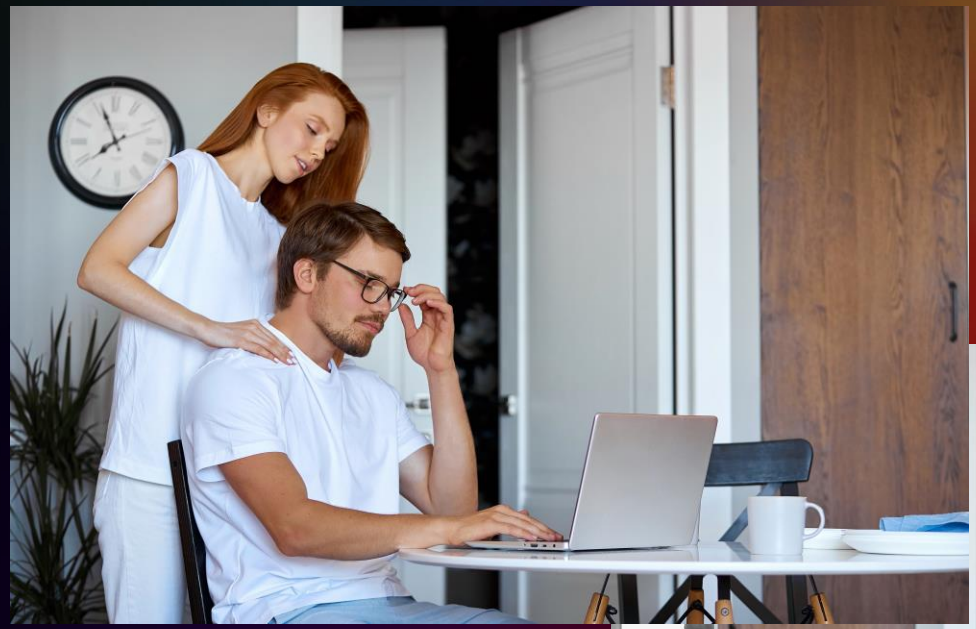
Personalized



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What do good human conversations sound like?

Contextually relevant





© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What do good human conversations sound like?

Relatable



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What do good human conversations sound like?

Pleasant, delightful



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What do good human conversations sound like?

Emotional intelligence and empathy

Personalized

Contextual

Relatable

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Machine learning in text-to-speech

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

TTS in Alexa



Utterance

"Play" directive

TTS

Alexa voice service

Recognize

Recognition result

Automatic speech recognition (ASR)

Recognition result

Intent

Natural language understanding (NLU)

Intent

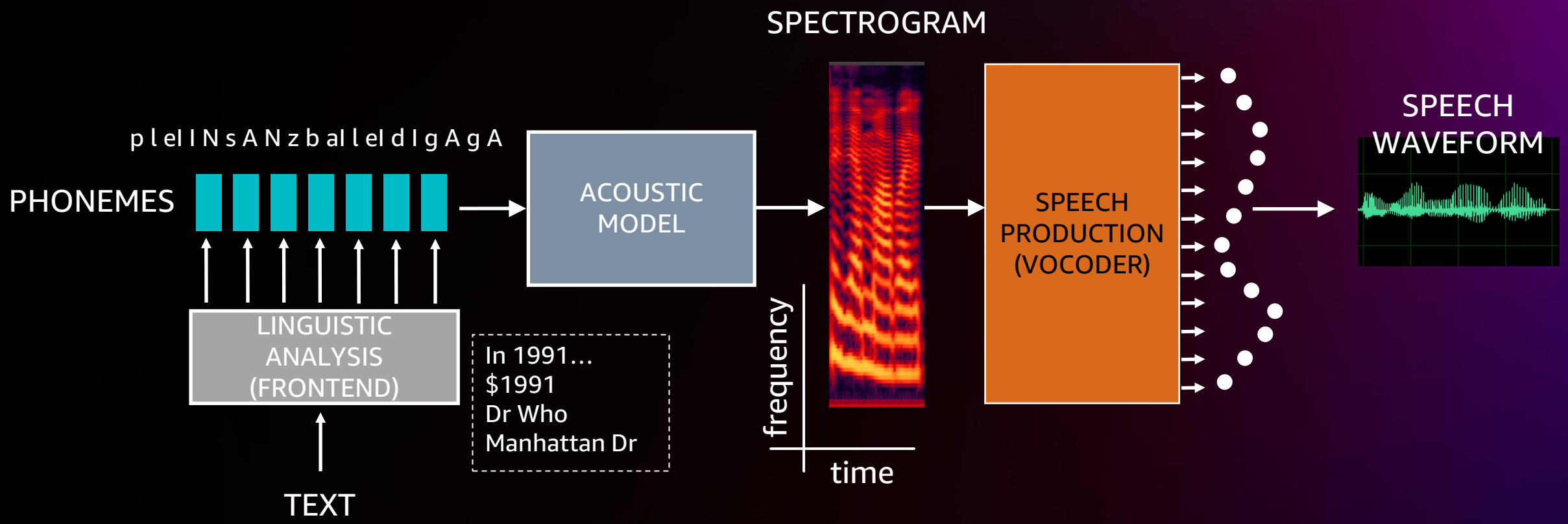
"Play" directive

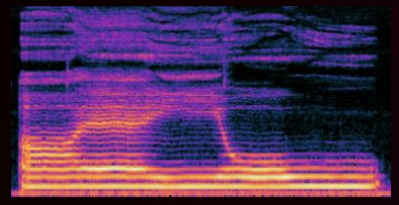
Skills

aws

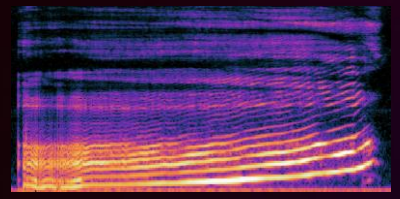
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ML-powered TTS





“a-e-i-o-u-”

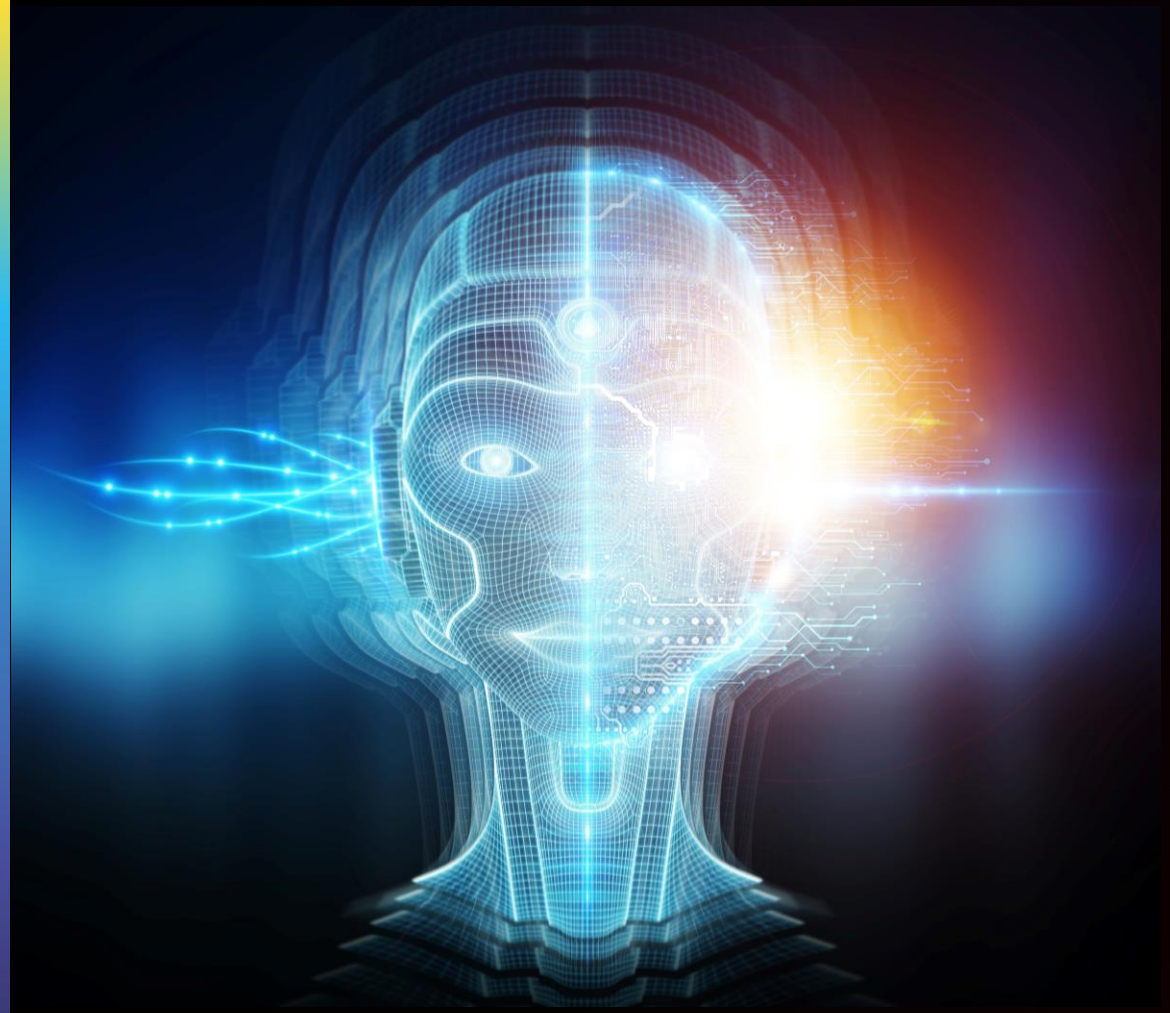


rising pitch

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Benefits of the ML approach



- More natural
- More flexible voices
- More scalable to new voices

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Machine learning for more human-like speech

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Emotional intelligence and empathy

Personalized

Contextual

Relatable

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Personalized speech

- What does personalized speech mean?
- A recognized child asks Alexa, “What is war?”

➢ Adult text, adult voice

➢ Kid-friendly text, teacher voice



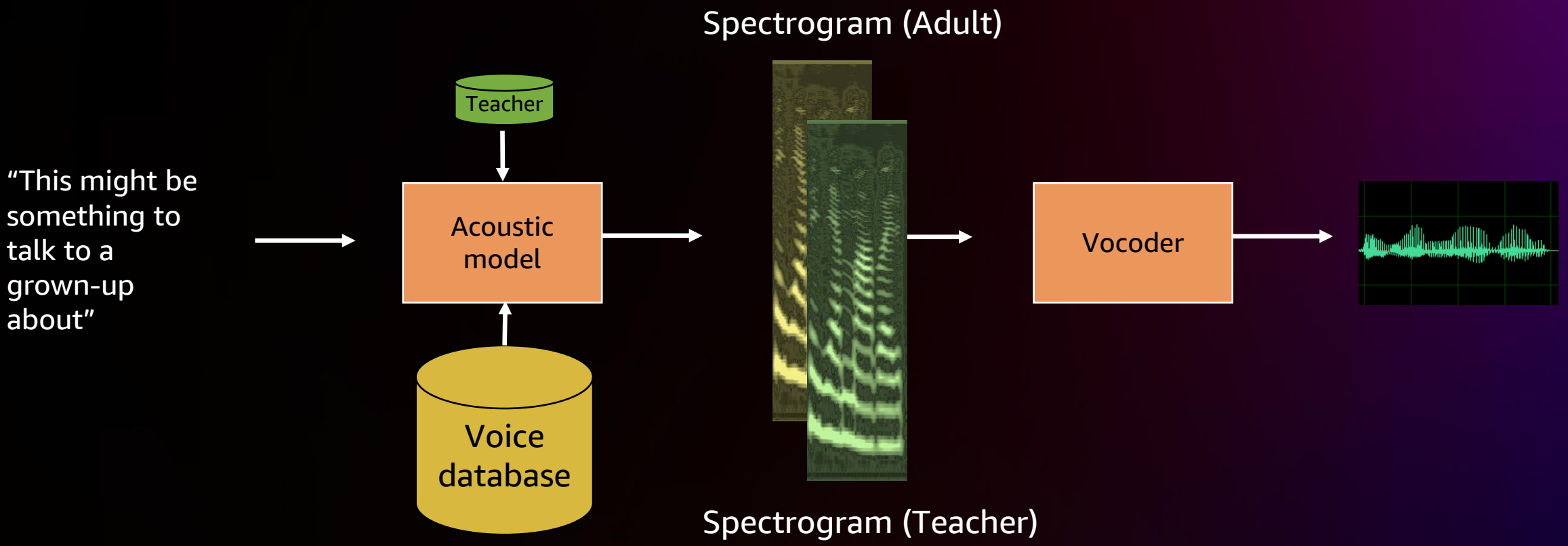


aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Personalized TTS: Role of ML

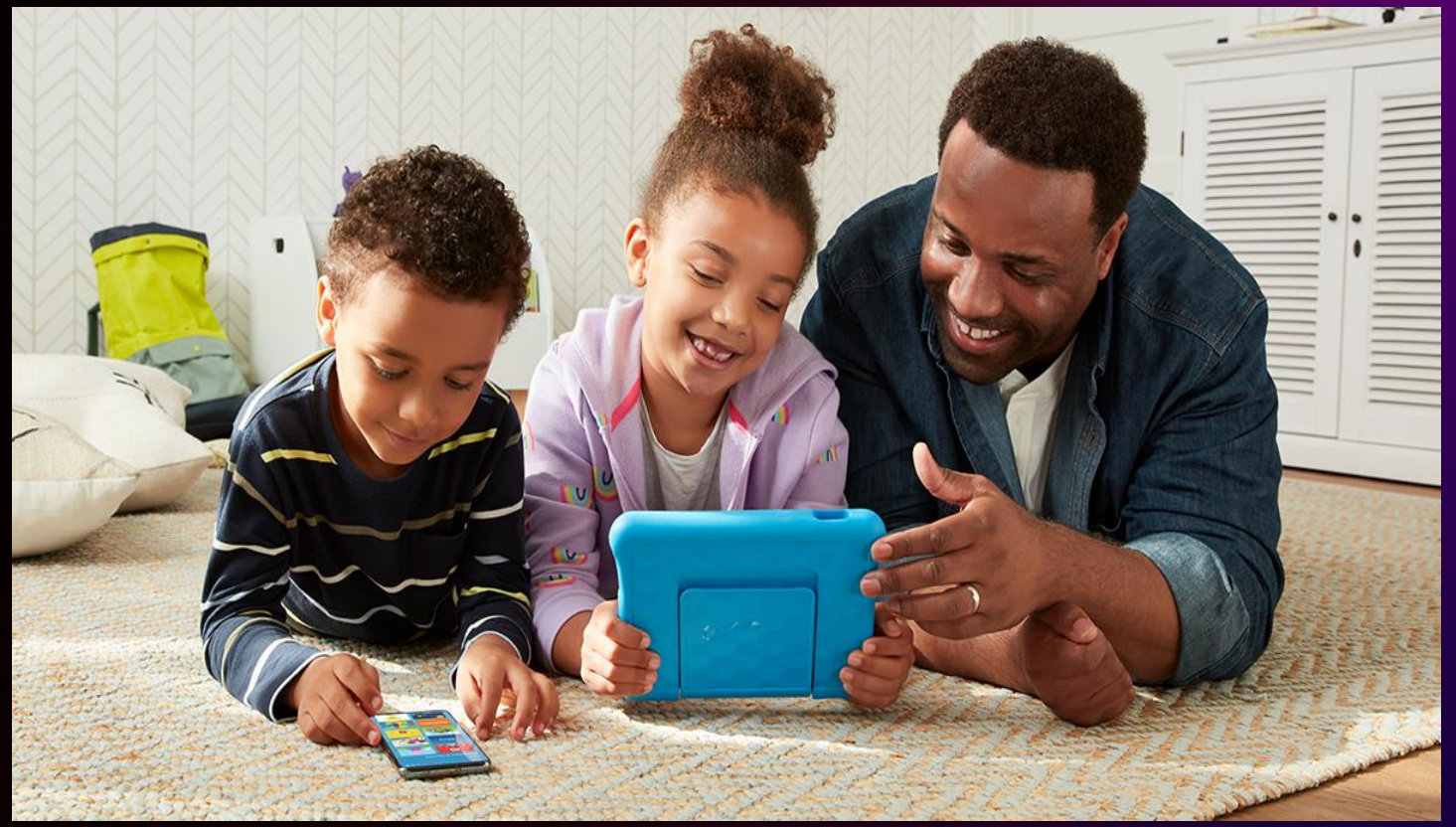
We can create new styles efficiently by supplementing an acoustic model with small amounts of extra data



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Benefits of personalized speech

- Trust, customer delight, and engagement
- Kids learn more when practicing math problems with Amazon Math when Alexa uses the teacher voice
- Our surveys show that parents also prefer the voice



aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Emotional intelligence and empathy

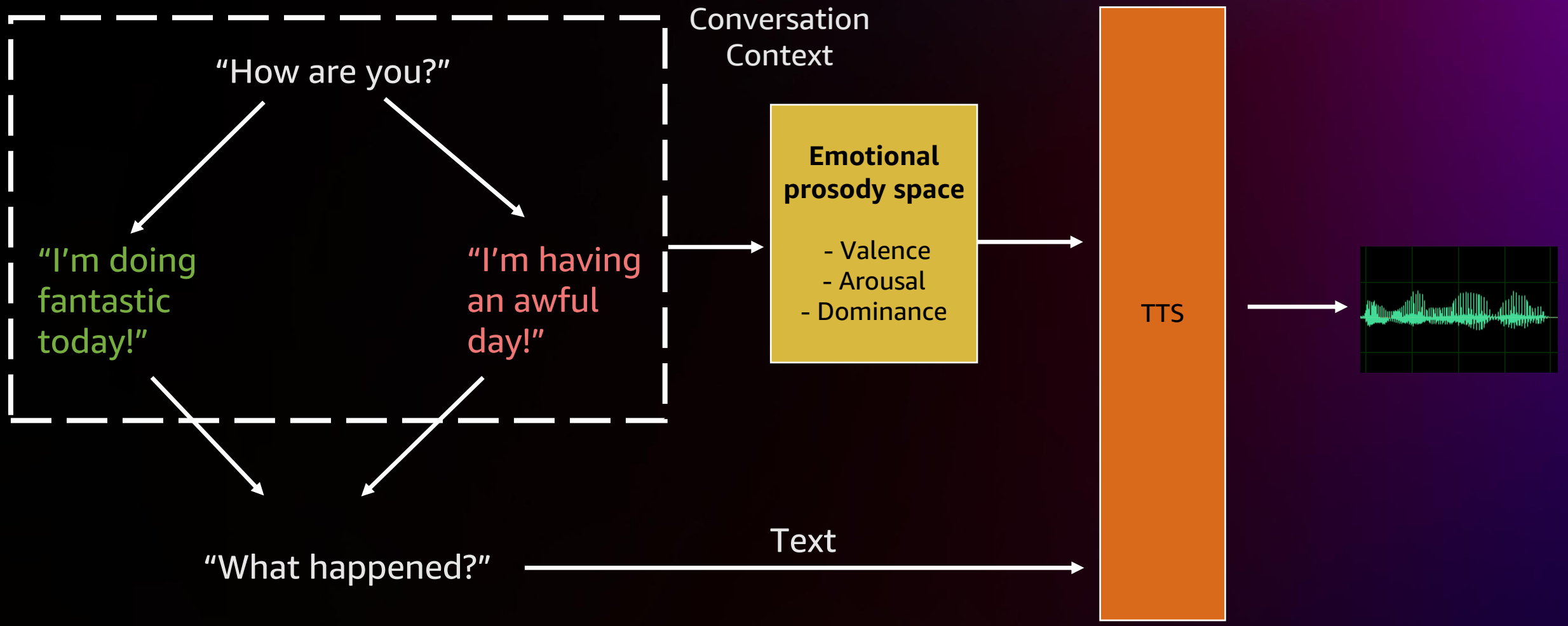
Personalized

Contextual

Relatable

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Conversation as context



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Adapting speech to the conversation

Alexa, help me find some music.

Happy to help. Let's find something you'll like on Amazon Music.
Any particular type of music you wanted?

Let's have some rock music.

Got it. How about something by The Eagles?

I'm not a big fan of The Eagles.

Hmm, okay. How about we try some songs by Queen?

No, give me something more modern!

Oh dear, I'm not getting you at all today.
What would you think about something by Foo Fighters?

That's more like it.

Great. In that case, do you think you'd like Vampire Weekend?

Now you're getting it . . .

Fantastic. Pulling a playlist together for you right now.

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

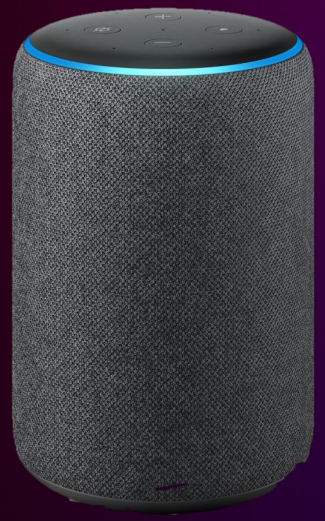
Benefits of contextual speech



Customers were more likely to listen to the recommended music



More likely to ask for more information about a product



Engaged with conversational AI more often and for longer

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Emotional intelligence and empathy

Personalized

Contextual

Relatable

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is speech disentanglement?

Speech “contains” a myriad of information



The aim is to separate these different aspects of speech

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

More relatable speech

- We want interacting with Alexa to be like talking to a friend or a family member
- We need a diverse set of voices to match Alexa's global customer base
- Improvements in:
  - Voice diversity
  - Language and accent diversity





Source: Neural architecture underlying person perception from in-group and out-group voices. Jiang, Sanford, and Pell, 2018, NeuroImage.

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

More relatable speech

- Two ML approaches can help us achieve this at scale:
  - **Transfer learning**
    Applying knowledge from one voice to another voice
  - **Zero-shot learning**
    Creating voices without any audio data





© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Accented voices

Not everyone speaks the Queen’s English



British English Alexa
speaking
Received Pronunciation (RP)



British English voice
speaking
Irish English

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

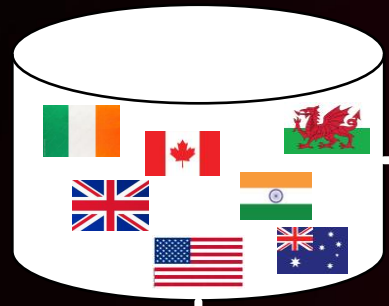
Accented voices

Not everyone speaks the Queen's English

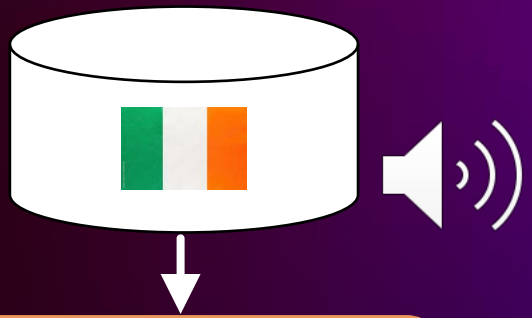


British English Alexa speaking Received Pronunciation (RP)

Recordings of multiple speakers of multiple English accents



Recordings of an Irish English speaker speaking Irish English



Voice conversion



Generate synthetic corpus of British English voice speaking Irish English

Polyglot

Speaker ID, accent ID

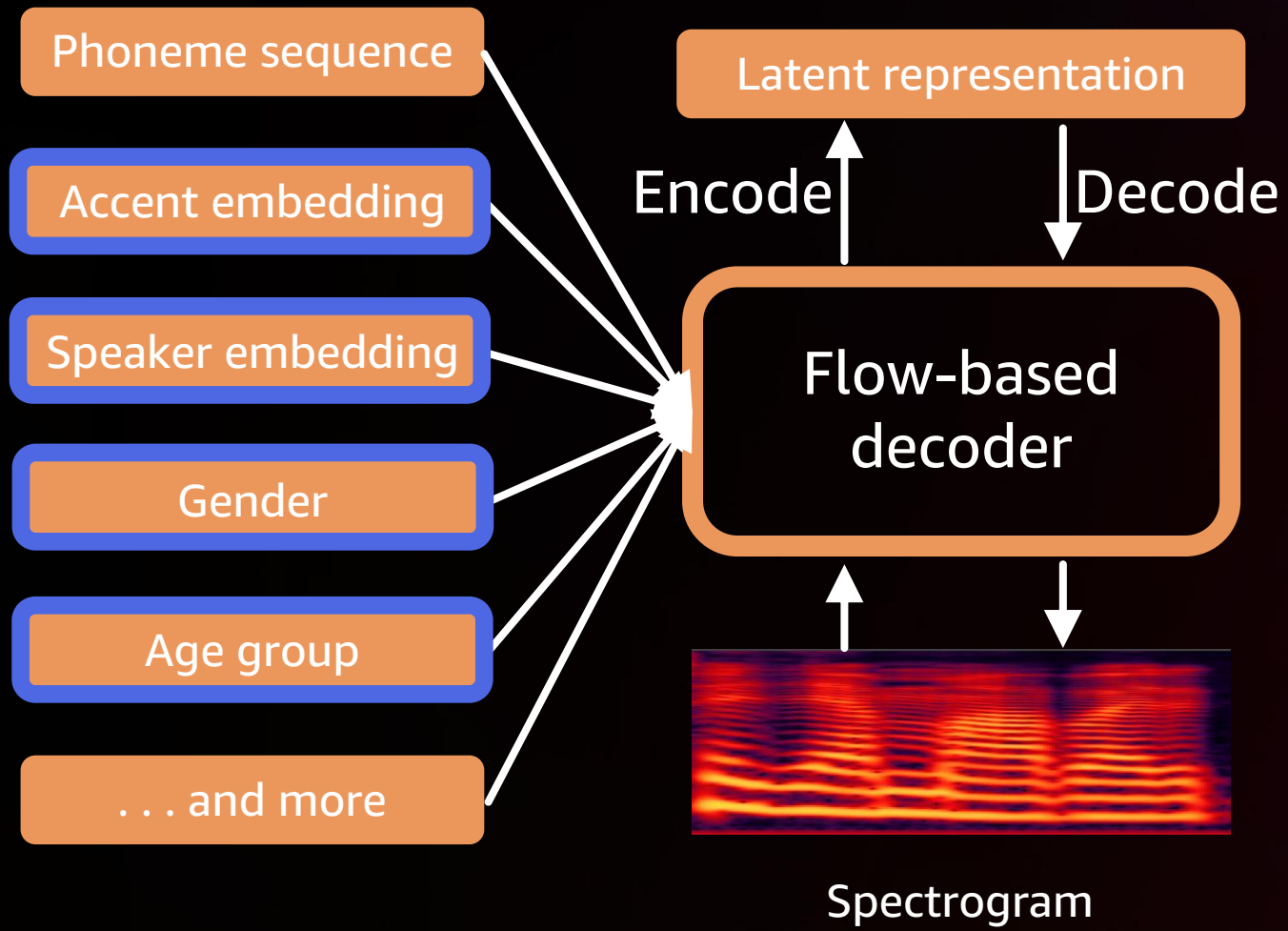


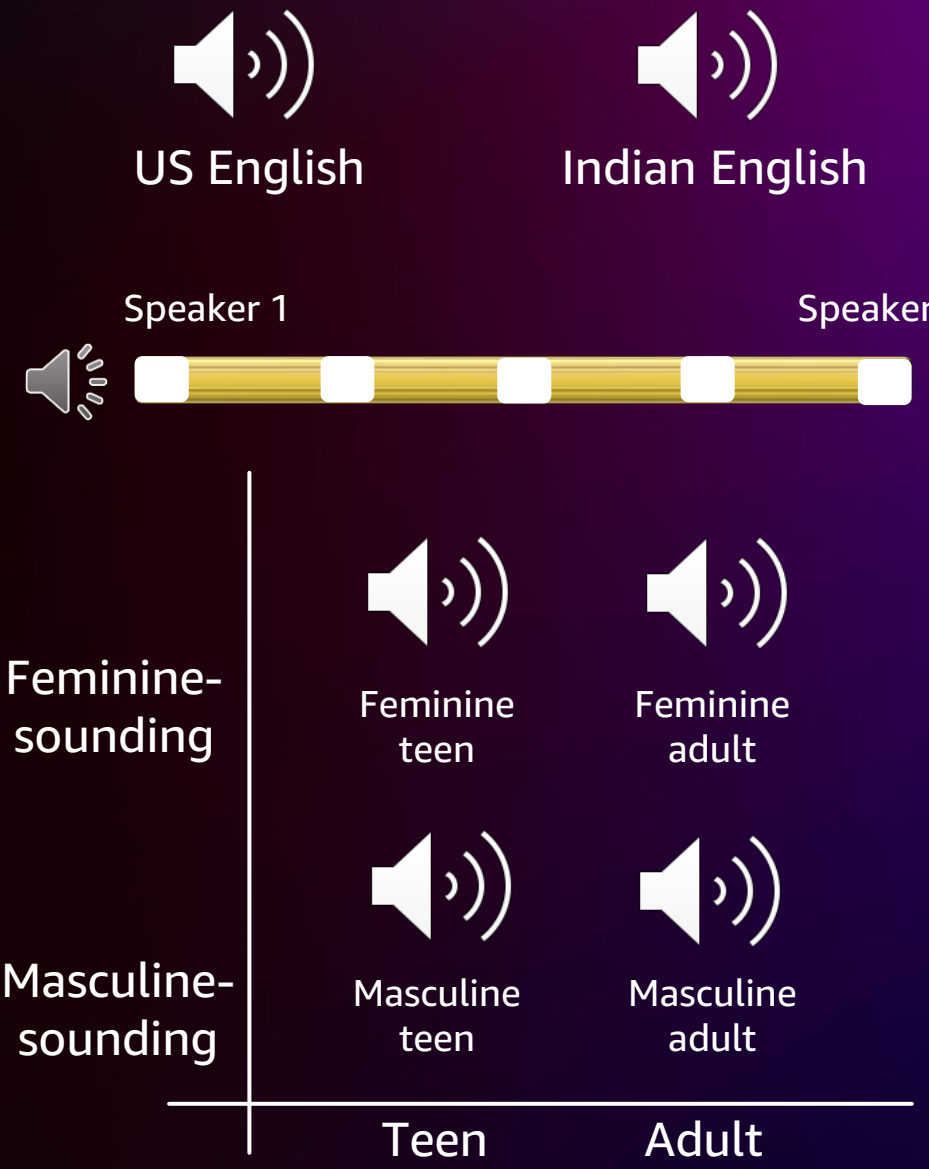
British English voice speaking Irish English

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Speech attribute control: Zero-shot learning





aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Benefits of human-like speech with Alexa



Productivity and convenience



Instantly available

Trusted advisor

Companion



Entertainment and fun



Information and knowledge

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Nikhil Minotra
minotran@amazon.com

Trevor Wood
trevowoo@amazon.com



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.