Microsoft

ELASTACLOUD

SPEAKING TO AZURE

Richard Conway, richard@elastacloud.com

Agenda

- History of speech in Computing
- Speech on Azure
- Video and audio extraction demos

Speech synthesis over the year (pre-Millenium)

1779: **Wolfgang von Kempelen** creates the first known speech synthesis machine, a mechanical device that could simulate simple speech sounds.

1939: **Homer Dudley** of Bell Labs introduces the **Voder**, the first electronic speech synthesizer, at the New York World's Fair.

1950s: **Dudley's Vocoder** (Voice Encoder), another Bell Labs invention, is developed and used for speech compression and transformation.

1961: **John Larry Kelly, Jr.** at Bell Labs uses an IBM 704 to create one of the first computer-generated voices to sing the song "Daisy Bell" ("Bicycle Built for Two").

1980s: The development of **DECtalk**, a speech synthesis system by **Digital Equipment Corporation**, becomes notable for its use in assistive technologies (e.g., used by Stephen Hawking).

1987: **AT&T Bell Labs** releases **Lucent Text-to-Speech (TTS)** system, which is one of the early commercially available text-to-speech systems.

1997: **Microsoft SAPI (Speech Application Programming Interface)** 4.0 is released, allowing developers to create speech-enabled applications for Windows.

Famous voices:

Sam – Microsoft SAPI 5

🔊

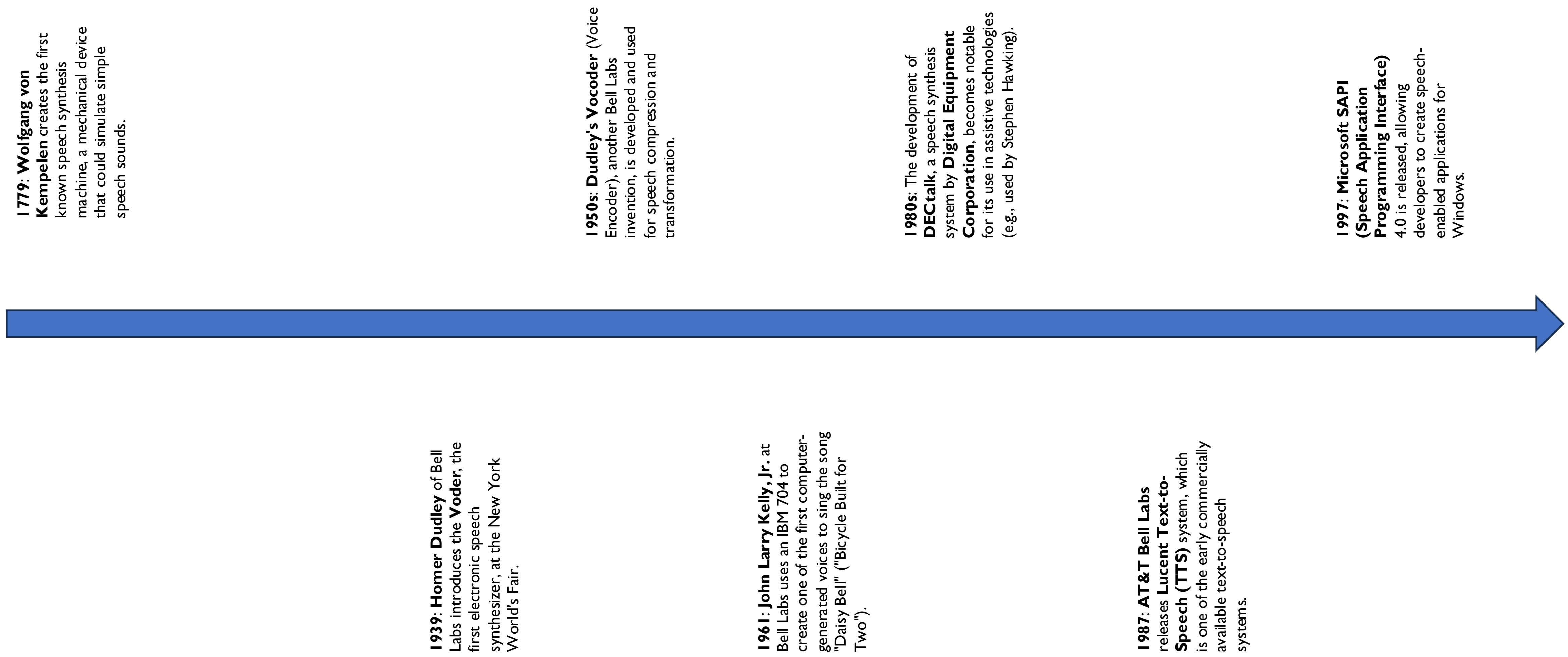
Stephen Hawking ([Stephen Hawking's Voice Emulator Project | Pawel Wozniak (pawozniak.com)](pawozniak.com))

🔊

Richard Conway

🔊

Speech synthesis over the year (> 2000)



1779: **Wolfgang von Kempelen** creates the first known speech synthesis machine, a mechanical device that could simulate simple speech sounds.

1939: **Homer Dudley** of Bell Labs introduces the **Voder**, the first electronic speech synthesizer, at the New York World's Fair.

1950s: **Dudley's Vocoder** (Voice Encoder), another Bell Labs invention, is developed and used for speech compression and transformation.

1961: **John Larry Kelly, Jr.** at Bell Labs uses an IBM 704 to create one of the first computer-generated voices to sing the song "Daisy Bell" ("Bicycle Built for Two").

1980s: The development of **DECtalk**, a speech synthesis system by **Digital Equipment Corporation**, becomes notable for its use in assistive technologies (e.g., used by Stephen Hawking).

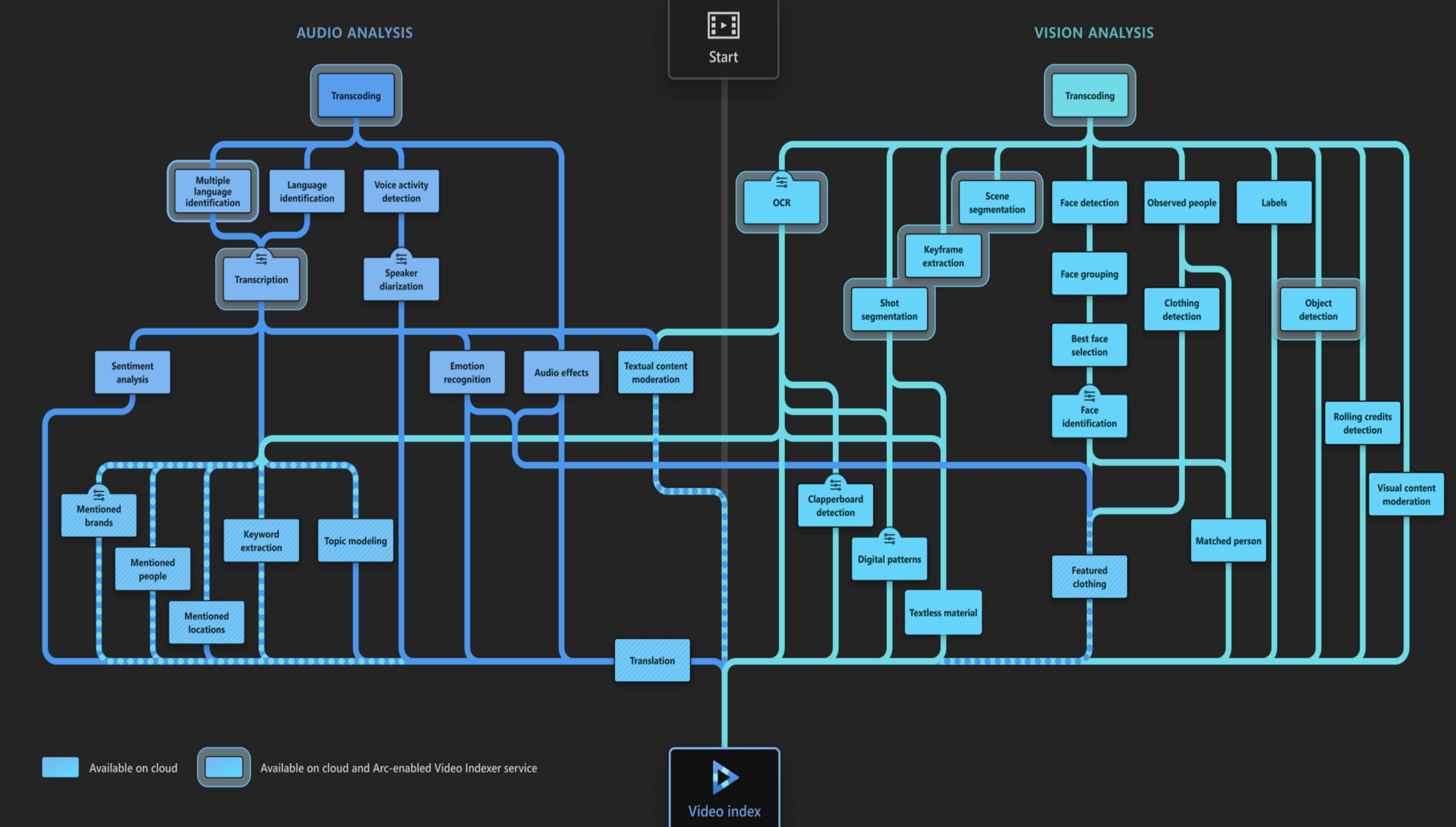
1987: **AT&T Bell Labs** releases **Lucent Text-to-Speech (TTS)** system, which is one of the early commercially available text-to-speech systems.

1997: **Microsoft SAPI (Speech Application Programming Interface)** 4.0 is released, allowing developers to create speech-enabled applications for Windows.

Resources you'll use

Azure provides a whole set of resources that you can use to build in Speech into your application

- Speech SDK
- Whisper
- AI Services
- Azure AI Video Indexer
- Speaker Recognition


Start
AUDIO ANALYSIS
VISION ANALYSIS
Transcoding
Multiple language identification
Language identification
Voice activity detection
Transcription
Speaker diarization
Sentiment analysis
Emotion recognition
Audio effects
Textual content moderation
Mentioned brands
Mentioned people
Mentioned locations
Keyword extraction
Topic modeling
Translation
Transcoding
OCR
Scene segmentation
Keyframe extraction
Shot segmentation
Face detection
Face grouping
Best face selection
Face identification
Observed people
Clothing detection
Labels
Object detection
Rolling credits detection
Visual content moderation
Clapperboard detection
Digital patterns
Textless material
Featured clothing
Matched person
Video index
Available on cloud
Available on cloud and Arc-enabled Video Indexer service

Recording Audio

- Use **pyaudio** library to record audio
- Sampling up to 44100 Hz
- Write 128 bps across 2 channels
- Set quality using discrete scale
- Use standard python streams
- **Frames** are sampled and must appended to **stream**

```
audio_format = pyaudio.paInt16
encoder = lameenc.Encoder()
encoder.set_bit_rate(128)
encoder.set_in_sample_rate(44100)
encoder.set_channels(1)
encoder.set_quality(2) # 2-high 5-medium 7-low

p = pyaudio.PyAudio() # Create a PyAudio session

# Open the microphone stream
stream = p.open(format=audio_format,
channels=channels,rate=sample_rate,
input=True, frames_per_buffer=1024)
```

Encapsulating voice

- Use SSML to define voice
- Can contain content and characteristics
- Can contain many voices
- Contains different voice roles
- Define whether voice is happy, sad, angry, whispering etc.

```
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
xml:lang="en-US">
   <voice name="en-US-AvaMultilingualNeural">
      Good morning!
   </voice>
   <voice name="en-US-AndrewMultilingualNeural">
      Good morning to you too Ava!
   </voice>
</speak>
```

```
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
xml:lang="en-US">
 <voice name="en-US-AvaMultilingualNeural">
   <mstts:express-as role=”YoungAdultFemale” style=“calm/angry …”>
      Good morning!
   </mstts:express-as>
</speak>
```

Using custom models

- Speech services allows you to train custom models
- Use models to provide the following:
  - Specialised vocab or domain specific terms for text to speech
  - Understand better accents and dialects (e.g. Scottish accent)
  - Cut out noise in noisier environments through better ”noisy” training set
  - Build in custom speech commands for security, home automation etc.
  - Speaker authentication and voice identification
  - Text to speech with custom voices
  - Multilingual support
  - Custom neural voice lite - Speech service - Azure AI services |

Speech Studio

- Use SSML to define voice
- Can contain content and characteristics
- Can contain many voices
- Contains different voice roles
- Define whether voice is happy, sad, angry, whispering etc.

```
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
xml:lang="en-US">
    <voice name="en-US-AvaMultilingualNeural">
        Good morning!
    </voice>
    <voice name="en-US-AndrewMultilingualNeural">
        Good morning to you too Ava!
    </voice>
</speak>
```

```
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
xml:lang="en-US">
 <voice name="en-US-AvaMultilingualNeural">
    <mstts:express-as role=”YoungAdultFemale” style=“calm/angry …”>
        Good morning!
    </mstts:express-as>
</speak>
```

Speech CLI

- Download and install spx
- Run voice tests from command line
- Customise voice using SSML

```
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
xml:lang="en-US">
    <voice name="en-US-AvaMultilingualNeural">
        Good morning!
    </voice>
    <voice name="en-US-AndrewMultilingualNeural">
        Good morning to you too Ava!
    </voice>
</speak>
```

```
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
xml:lang="en-US">
 <voice name="en-US-AvaMultilingualNeural">
    <mstts:express-as role=”YoungAdultFemale” style=“calm/angry …”>
        Good morning!
    </mstts:express-as>
</speak>
```

ELASTACLOUD

Microsoft

DEMO: Recording Audio

ELASTACLOUD

Microsoft

DEMO: Voice of Azure

ELASTACLOUD

Microsoft

DEMO: CLI + Custom voices

ELASTACLOUD

Microsoft

DEMO: Transcription

ELASTACLOUD

Microsoft

DEMO: Video decomposition

ELASTACLOUD

Microsoft

DEMO: Video Indexing

ELASTACLOUD

Microsoft

DEMO: Voice Enrolment

ELASTACLOUD

Microsoft

DEMO: Interview Mode

ELASTACLOUD

Richard Conway
richard@elastacloud.com

Elastacloud Limited
Spitalfields Works, 11 Toynbee Street, London, E1 7NE
Company Number: 07900393