AWS

SUMMIT

Amazon Polly

A service that turns text into lifelike speech

Remus Mois – Amazon Text-to-Speech

May 18, 2017

amazon
web services

What to Expect from the Session

- Introduction to Amazon Polly
- Features and functionalities
- Text-to-Speech: Under the Hood
- Getting started
- Pricing
- Introducing Vicki

Introduction to Amazon Polly

amazon webservices

Why we built Polly

- Apps using voice to communicate with end-users are becoming more common every day
- Naturalness of generated speech is a key element of user experience
- Integration of speech varies across use cases

What is Polly

- A service that converts text into lifelike speech
- Offers 48 lifelike voices and 24 languages
- Low latency responses enable developers to build real-time systems
- Developers can store, replay and distribute generated speech

Polly – Quality

Natural sounding speech

A subjective measure of how close TTS output is to human speech.

Accurate text processing

Able to interpret common text formats such as abbreviations, numerical sequences, homographs, etc.

“In Berlin ist es heute sonnig, bei Temperaturen von 5 bis 10°C.”

“Im Anhang finden Sie Bl. 16-17." → "Im Anhang finden Sie Blätter 16 bis 17.”

“Das Radio spielt Radio Gaga von Queen.”

Highly intelligibile

A measure of comprehensibility of speech.

“Fischers Fritze fischt frische Fische.”

Polly – Language Portfolio

EMEA:

- Danish
- Dutch
- British English
- French
- German
- Icelandic
- Italian
- Norwegian
- Polish
- Portuguese
- Romanian
- Russian
- Spanish
- Swedish
- Turkish
- Welsh
- Welsh English

Americas:

- Brazilian Portuguese
- Canadian French
- English (US)
- Spanish (US)

A-PAC:

- Australian English
- Indian English
- Japanese

Features and Functionality

Polly features: SSML

Speech Synthesis Markup Language

A W3C recommendation, an XML-based markup language for speech synthesis applications

```
<speak>
    Ich heiße <phoneme alphabet='x-sampa' ph='"RE.mUs "mOYs'/>.
    Meinen Nachnamen schreibt man so:
    <prosody rate='x-slow'>
        <say-as interpret-as="characters">Mois</say-as>
    </prosody>
</speak>
```

Whispering effect

- Mimics effect of whispering
- Effect can be applied to any Amazon Polly voice via SSML

```
<speak>Ich werde euch mein Geheimnis erzählen.
<amazon:effect name="whispered">Bielefeld gibt es gar
nicht.</amazon:effect></speak>
```

Polly features: Lexicons

Enables developers to customize the pronunciation of words or phrases

Meine Tochter heißt Alice. 🔊 🔊

```
<lexeme>
    <grapheme>Alice</grapheme>
    <grapheme>alice</grapheme>
    <grapheme>ALICE</grapheme>
    <phoneme>”?E.lIs</phoneme>
</lexeme>
```

https://aws.amazon.com/documentation/polly/

Speech Marks

- Enable developers to synchronize speech with visual animation
- Enable developers to implement karaoke-style word highlighting
- Speech Marks are a stream of metadata with information about offsets of specific sentence, word, phoneme (visemes)

Text-to-Speech: Under the Hood

Main Challenges of Text-to-Speech

Goal: Convert text into intelligible, accurate, and natural speech

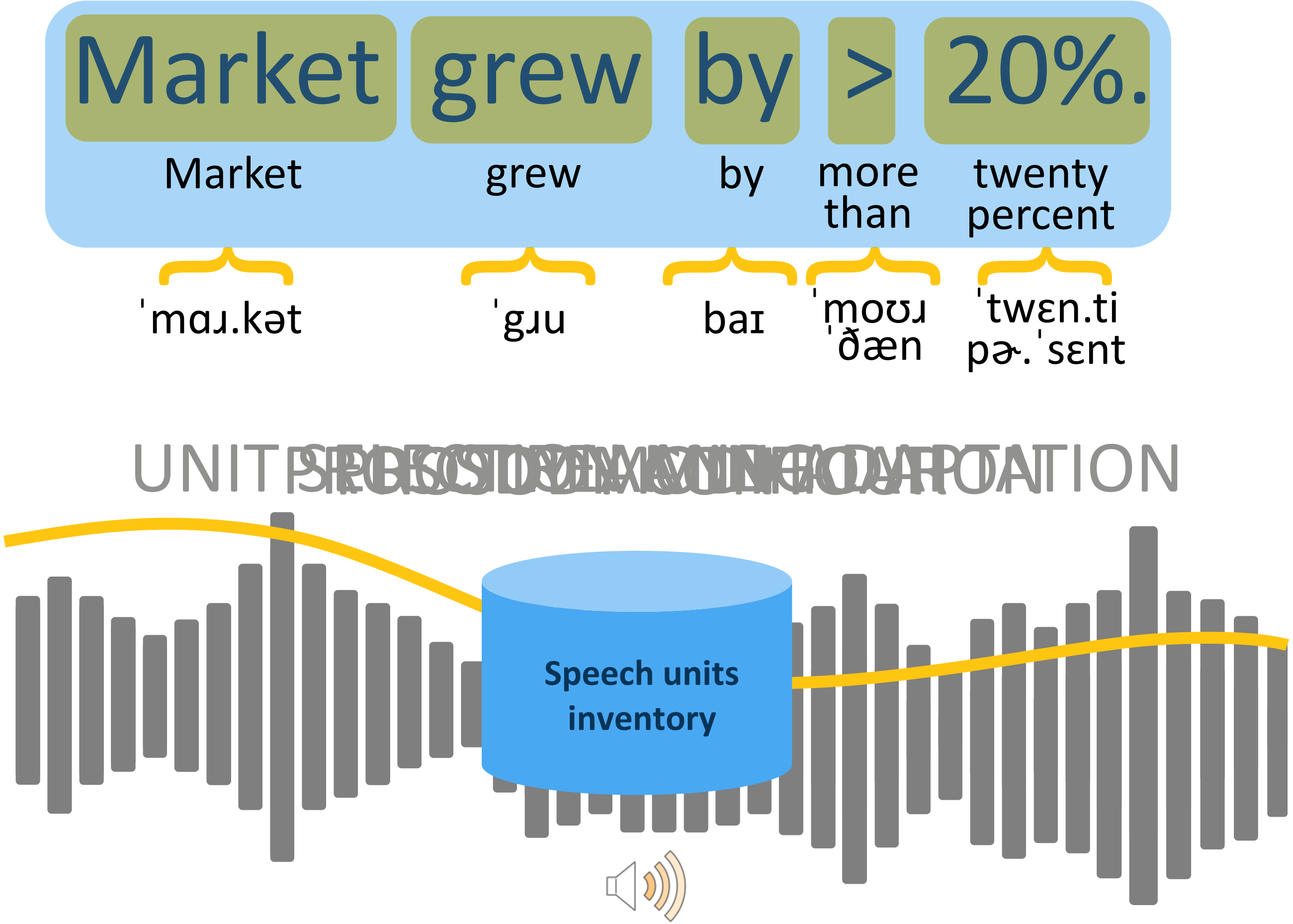
Challenges

- Homographs: words written identically that have different pronunciation

  *Die **Band** spielt heute Abend. vs Die Aufnahme kommt vom **Band**.*

- Text normalization: *‘**a.d.**’ expanded as ‘**an der**’ in address contexts (**Neustadt a. d. Weinstraße**)*
- Correct grammatical case for units (***nach 3l***), the number 1 (***1 Stunde, in 1 Stunde***) and …
- Foreign words (**enfant terrible**), proper names (***Seven Nation Army***) etc.

TEXT PROCESSING / WORDS / PHONEMES



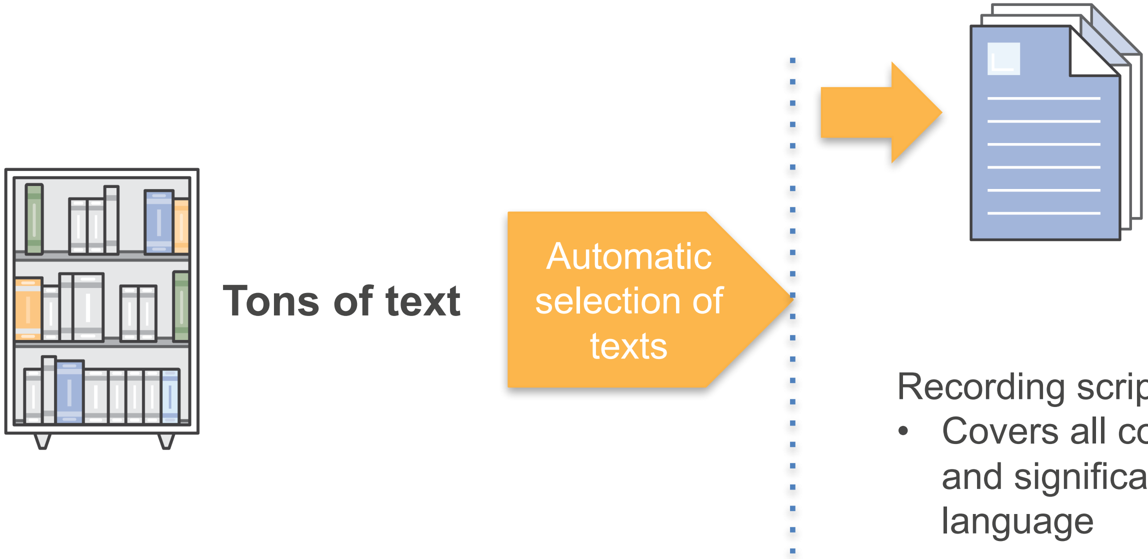
Market grew by > 20%.

Market grew by more than twenty percent

ˈmɑɹ.kət ˈgɹu baɪ ˈmoʊɹ ˈðæn ˈtwɛn.ti pɚ.ˈsɛnt

UNIT SELECTION / PROSODY MODIFICATION / CONCATENATION / SPEECH

Recording Data for TTS



Tons of text

Automatic selection of texts

Recording script:
Few weeks of recordings

Recording script:

- Covers all combinations of diphones and significant features in a language

An apple a day, keeps …

an error occurred while searching for your route

because s**nap**s weren't all so obedient anymore,

now we say ap**ple a**gain. and we say apple,

general electric soars t**oday.** information on general electric

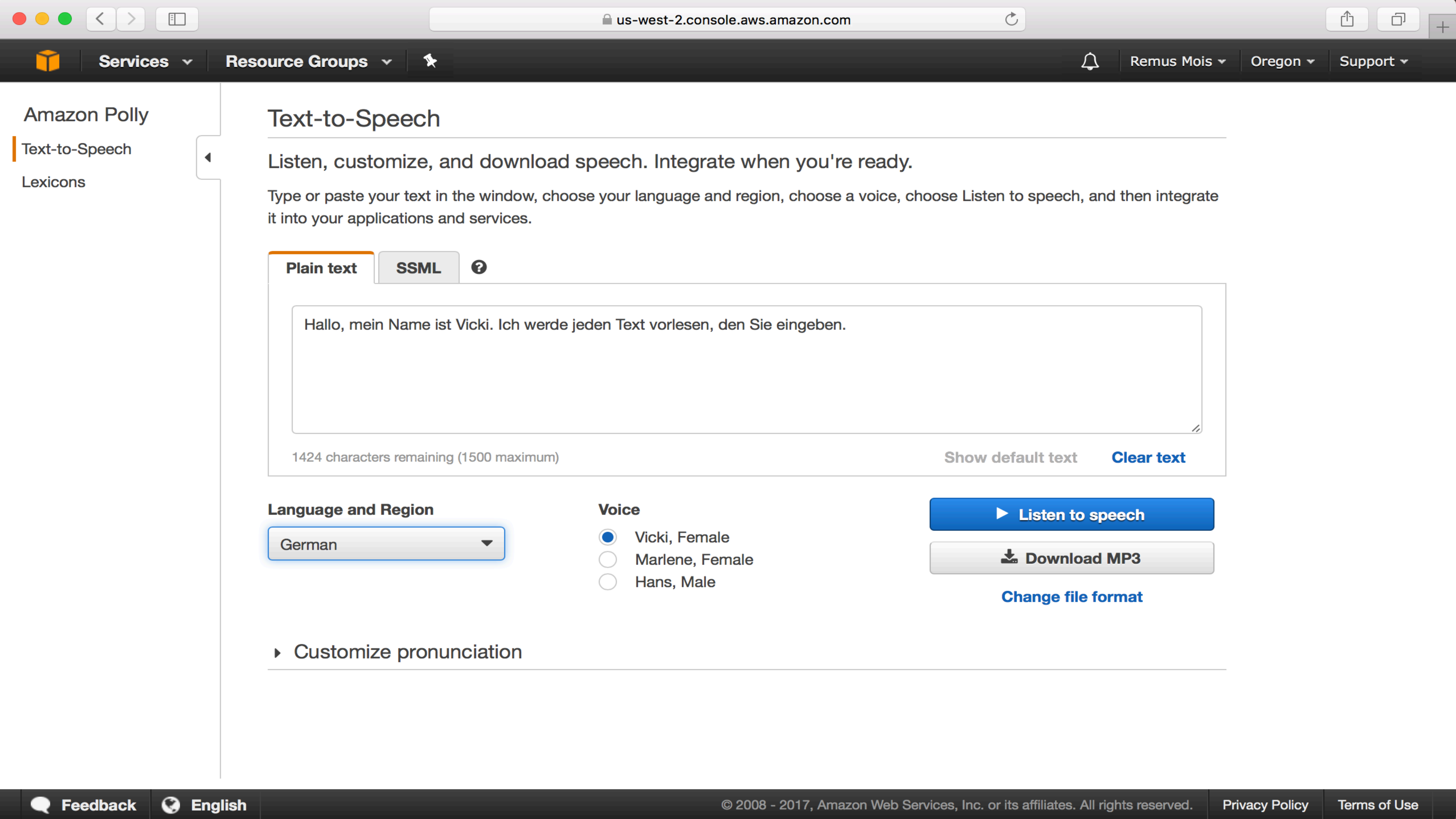
quick breads, zucchini, holiday**, c**rock pot, cake,

so are you still **keep**ing tabs on your old team,

that weighs more than four tons, disru**pts** the herring's swim

…

Getting started



Text-to-Speech

Listen, customize, and download speech. Integrate when you're ready.

Type or paste your text in the window, choose your language and region, choose a voice, choose Listen to speech, and then integrate it into your applications and services.

Hallo, mein Name ist Vicki. Ich werde jeden Text vorlesen, den Sie eingeben.

1424 characters remaining (1500 maximum)

Language and Region

German

Voice

- ● Vicki, Female
- ○ Marlene, Female
- ○ Hans, Male

© 2008 - 2017, Amazon Web Services, Inc. or its affiliates. All rights reserved.

First app

```
from boto3 import Session
from contextlib import closing

polly = Session().client("polly")

response = polly.synthesize_speech(
    Text="Hallo Welt!",
    OutputFormat="mp3",
    VoiceId=”Vicki")

with closing(response["AudioStream"]) as stream:
    with open("speech.mp3", "wb") as file:
        file.write(stream.read())
```

Polly is cost-effective

- Pay-as-you-go
- $4 for 1M characters
- Free Tier of 5M characters/month - first year
- You can store and reuse generated speech

Introducing Vicki

Introducing Vicki – the new Polly German voice

- Wide coverage of speech units resulting in higher naturalness
- Support for Denglisch (frequent Anglicisms in German)

AWS

SUMMIT

amazon
web services