

# YOUR VOICE IS MY PASSPORT

John Seymour and Azeem Aqil

# Introduction

Voice is starting to be used for authentication

Machine Learning Journal

## Personalized Hey Siri

Vol. 1, Issue 9 • April 2018  
by Siri Team

Apple introduced the “Hey Siri” feature with the iPhone 6 (iOS 8). This feature allows users to invoke Siri without having to press the home button. When a user says, “Hey Siri, how is the weather today?” the phone wakes up upon hearing “Hey Siri” and processes the rest of the utterance as a Siri request.

The feature’s ability to listen continuously for the “Hey Siri” trigger phrase lets users access Siri in situations where their hands might be otherwise occupied, such as while driving or cooking, as well as in situations when their respective devices are not within arm’s reach. Imagine a scenario where a user is asking his or her iPhone 6 on

# Introduction

Voice is starting to be used for authentication

The screenshot shows a search results page from Google. The top bar includes the Apple logo and the text "Machine Learning Journal". The main search result is titled "Personalized Hey Siri" and is dated "Vol. 1, Issue 9 • April 2018 by Siri Team". Below the title, there's a snippet of text: "Apple introduce feature allows button. When phone wakes utterance as". To the right of the snippet, there's a sidebar with a search bar labeled "Describe your issue" and a "Google" logo. The main content area starts with "Step 3. Enjoy media or ask for personal information". It explains that once Voice Match is set up, users can enjoy media using voice commands and lists various topics: Routines (in U.S.), My Day (other countries), Calendars, Flights, Services, Shopping lists, Payments, and Photos. Below this, there's a section titled "Voice Match & feature functionality" with dropdown menus for "Media" and "Home control devices".

Apple introduce feature allows button. When phone wakes utterance as

Step 3. Enjoy media or ask for personal information

Once you set up Voice Match, you can [enjoy media using voice commands](#). You can also get quick answers to these personal topics:

- [Routines \(in U.S.\)](#), [My Day \(other countries\)](#)
- [Calendars](#)
- [Flights](#)
- [Services](#)
- [Shopping lists](#)
- [Payments](#)
- [Photos](#)

Voice Match & feature functionality

Media

Home control devices

# Introduction

Voice is starting to be used for authentication

Machine Learning Journal

## Personalized Hey Siri

Vol. 1, Issue 9 • April 2018  
by Siri Team

Google

Apple introduces feature allowing users to control their phone with voice commands. The feature allows users to access Siri in situ while driving or within arm's reach.

The feature's aim is to make it easier for users to interact with their phones without having to physically touch them. It can be triggered by saying "Hey Siri" or by using a specific wake word.

Once you set up Voice Match, you can enter these personal topics:

- Routines (in U.S.), My Day (other countries)
- Calendars
- Flights
- Services
- Shopping lists
- Payments
- Photos

### Step 3. Enjoy media on the go

Once you set up Voice Match, you can enter these personal topics:

- Routines (in U.S.), My Day (other countries)
- Calendars
- Flights
- Services
- Shopping lists
- Payments
- Photos

### Schwab voice ID Service

Access your account with your voice. Schwab's voice ID service allows you to access your account just by speaking one simple phrase, "At Schwab, my voice is my password."

\* Schwab's voice ID service is not available on all contact numbers.

No more personal questions. No more PINs.

Schwab is introducing a new voice ID service that uses voice biometrics technology to identify you by your unique voice. Whether you want to use our automated phone service or speak with one of our Financial Professionals, our voice ID service is one of the fastest and most convenient ways to securely identify yourself over the phone.

When you call us, you will simply be prompted to say the passphrase "At Schwab, my voice is my password" to be securely verified.

Need help? Call 800-435-4000.

#### Frequently Asked Questions

- Is my voiceprint really unique?  
Yes, just like your fingerprint, your voiceprint is uniquely yours.
- Will the voice ID service work if I have a cold?  
Yes, our voice ID service verifies hundreds of voice characteristics, only a few of which are affected by a cold.
- Does background noise affect performance?  
Yes, like with speech recognition, background noise can negatively impact performance. That's why our solution offers the leading technology that has the best noise-canceling algorithms in the industry. Clients should be able to use voice biometric authentication from a location with typical noise.

### Voice Match & feature functionality

Media

Home control devices

Print



# Introduction

Voice is starting to be used for authentication

The collage consists of several screenshots:

- Apple Machine Learning Journal:** A screenshot showing a news article about Siri's Personalized Hey Siri feature. It includes a snippet of text: "Apple introduces feature allowing users to wake Siri with a simple utterance as it's within arm's reach".
- Schwab voice ID Service:** A screenshot of a website page for Schwab's voice ID service. It explains that users can access their account by speaking a phrase like "At Schwab, my voice is my password".
- Microsoft Azure Speaker Recognition:** A screenshot of the Microsoft Azure Cognitive Services page for Speaker Recognition. It shows a preview image of two people and a button to "Try Speaker Recognition".
- Google Assistant:** A screenshot of the Google Assistant interface showing the "Step 3. Enjoy media" section. It lists various routines like Routines (in U.S.), My Day (other country), Calendars, Flights, Services, Shopping lists, Payments, and Photos.
- Voice Match & feature functionality:** A screenshot of a section titled "Voice Match & feature functionality" which includes "Media" and "Home control devices" sections.
- Frequently Asked Questions (FAQ):** A screenshot of a FAQ section for the voice ID service, addressing topics like voice uniqueness, colds, background noise, and performance.
- Speaker Verification:** A screenshot of a section titled "Speaker Verification" explaining how to use the API for verification, including enrollment steps and sample phrases.

# Goal: Break Voice Authentication (with minimal effort)

# Obligatory Sneakers Reference



# Obligatory Sneakers Reference



And here's how you do it.

- In Sneakers, they social engineered the target in order to record the exact words they needed
- In practice, this is hard to do
- Luckily, text-to-speech exists



# Overview of Text to Speech (TTS)

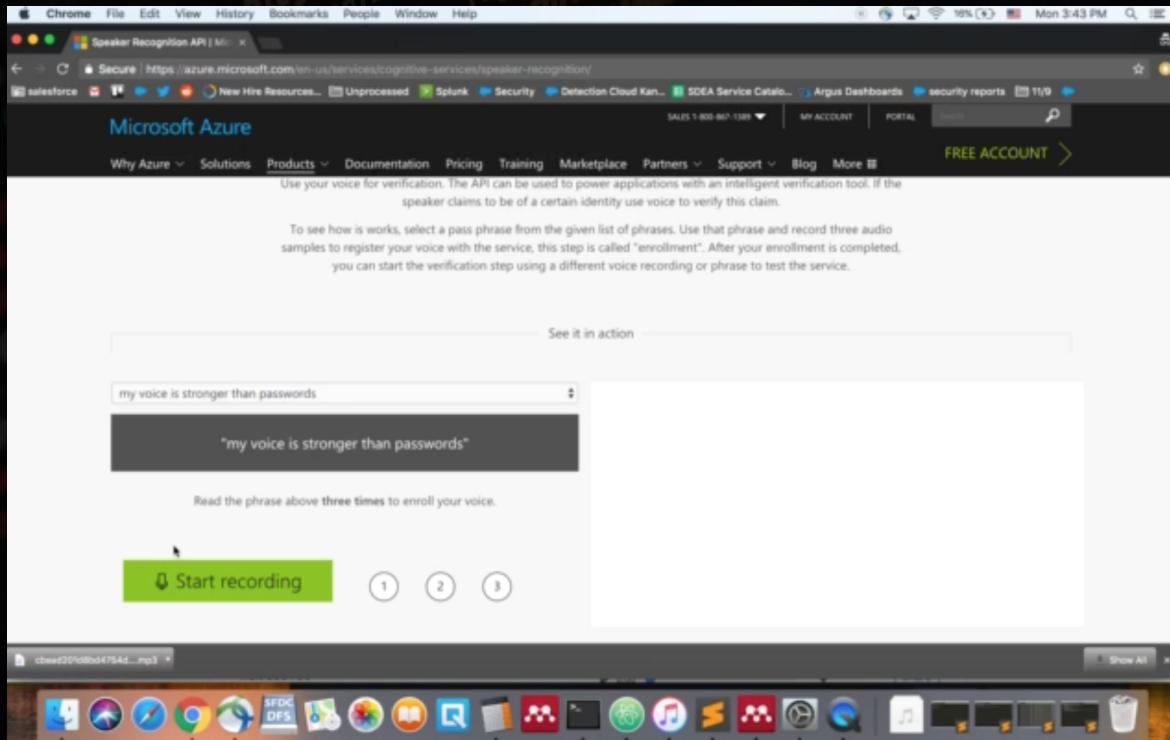
- Essential idea: you give the algorithm text, and it generates the equivalent audio representation of that text (e.g. Mel Spectrograms)
- Model learns the mapping between the transcript and audio
- The way it does this is you give it labeled (transcribed) audio and feed it into a deep neural network
  - Generally models are trained on a single person's voice
  - Generally deep learning models require a LARGE amount of labeled data
    - Open source datasets (e.g. Blizzard, LJ Speech)
    - >24 hours of labeled data to do well

# Proof of concept



- Create an account
- Record >30 sentences
  - Chosen by Lyrebird, same for all users.
- Provide a target sentence that Lyrebird will generate.
  
- Apple Siri and Microsoft Speaker Recognition API (Public Beta)
- Proof of concept is of limited value from a security standpoint

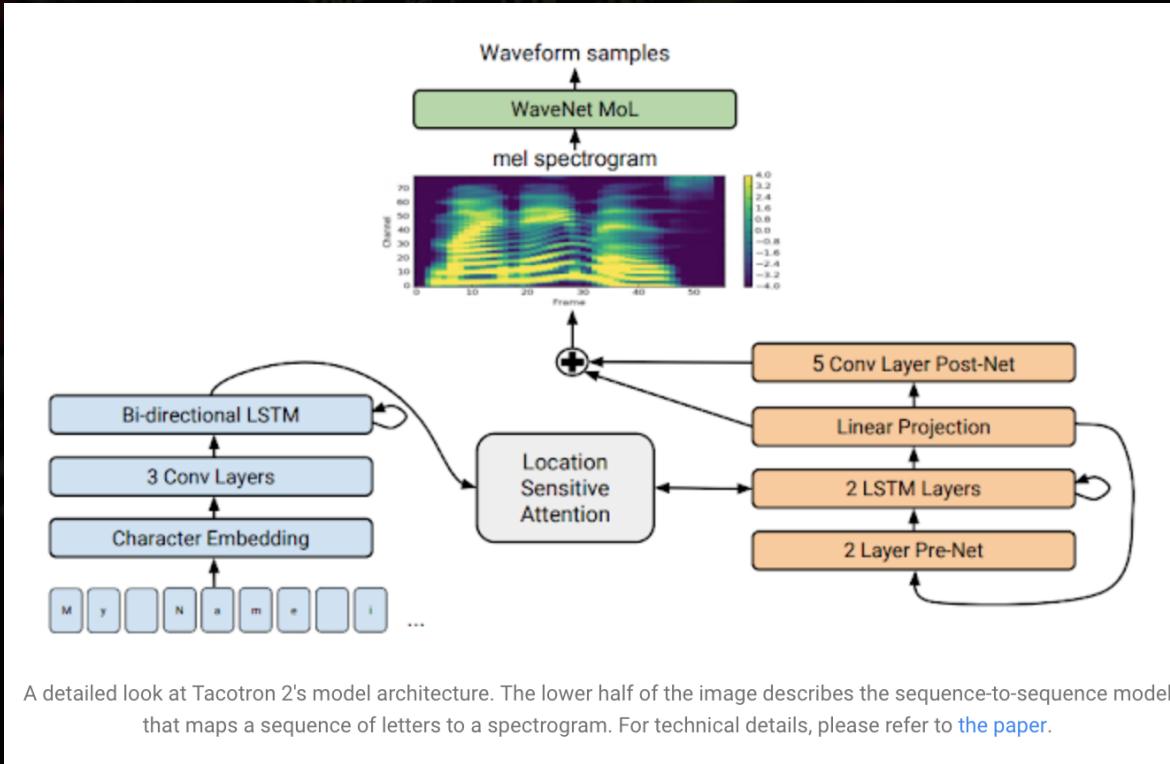
# Proof of Concept Demo



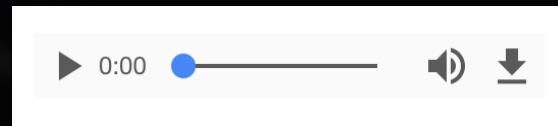
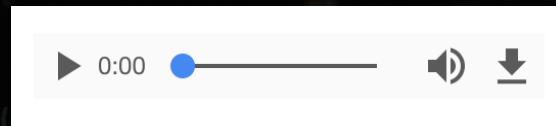
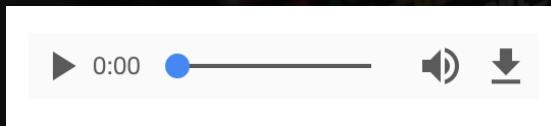
# Open source TTS models

- Several open source models (Tacotron, Wavenet are best known)
  - WaveNet generates realistic human sounding output, however, needs to be ‘tuned’ significantly.
- Tacotron simplifies this process greatly
  - The production of the feature set (which needs tuning in WaveNet) is replaced by another NN that works directly off data
- We use Tacotron

# Model overview



# Tacotron v1, Tacotron v2, and WaveNet



# Collecting Data

- That's all well and good, but in order to train a model, we need to feed it data
  - Can grab audio from e.g. YouTube, but quality/quantity are both important
- Need to transcribe this data
  - Youtube/Google Speech API not good enough
  - We also had to cut pieces with poor quality or with “um”
  - We had to manually transcribe the data we used for training
- Most open source models require short (<10 second) snippets of audio
  - Use ffmpeg to split the file into chunks

# Data Augmentation

- Publicly available data for a specific target is probably limited
- Transcribing is time intensive since it must be done manually
- Need LARGE amount of high quality training data
- Solution: Augment Data

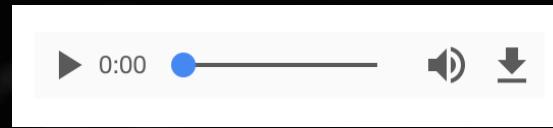
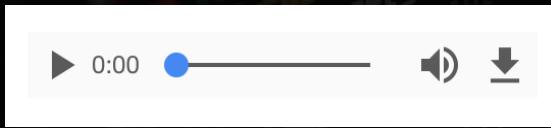
# Data Augmentation: Shifting Pitch

- Slow down and speed up audio to generate new examples
- Libraries (pydub) available for this
- Measured how far pitch can be raised/lowered by recording “Hey Siri” and testing how much speed up/slowdown would be accepted
  - 0.88x to 1.21x for our tests
  - YMMV for exact parameter (probably different for every person)

# Data Augmentation: Transfer Learning

1. Initially train on large open source dataset (Blizzard, LJSpeech)
2. Get a good model, stop training
3. Replace open source dataset with the target data
4. Continue to train

# Transfer Learning Demo



# Putting it all together

1. Scrape data from target (e.g. Youtube)
2. Select high-quality samples
3. Transcribe and chunk audio
4. Augment audio by shifting pitch
5. Train general TTS model on open source dataset
6. Replace general model training data with target data; finish training
7. Synthesize voice from model

# Our work in perspective (ML for offense)

- Attacks on ML systems
  - Adversarial Attacks
    - Most prior work attacking voice systems utilize GANs
    - Pro: hiding commands within benign-sounding audio
    - Con: method is currently brittle
    - (We use the simpler approach of generating speech for a given user)
  - Poisoning the Well
  - Privacy/Differential Privacy
- Attacks using ML systems
  - Phishing
  - DeepFakes
  - Robotics/Social Engineering

# Mitigation: MFA

- Defense in Depth
- Potential issue: Speaker Recognition with unknown vocabulary is hard
- Potential issue: Passphrases may not be kept secret

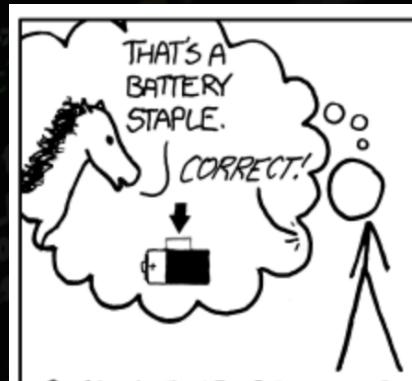


Image credit: XKCD.com

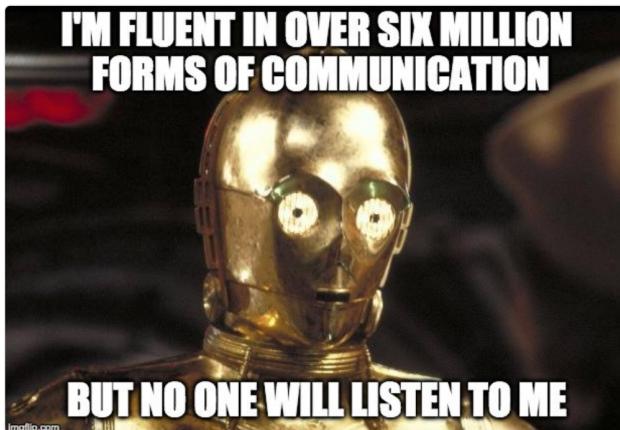
# Mitigation: Detect CGA



Patrick Traynor  
@patrickgtraynor

Following

We've developed a technique for protecting voice interfaces by identifying whether or not the source of the audio was a human being or a speaker (e.g., from your TV, smartphone, #iot device, etc). 1/ @uf\_fics @ufcise



5:19 AM - 18 Jun 2018

## MIT News

ON CAMPUS AND AROUND THE WORLD

Browse

or Search



FULL SCREEN



Arnav Kapur, a researcher in the Fluid Interfaces group at the MIT Media Lab, demonstrates the AlterEgo project.

Image: Lorrie Lejeune/MIT

Computer system transcribes words users “speak silently”

Electrodes on the face and jaw pick up otherwise undetectable neuromuscular signals triggered by internal verbalizations.

# Conclusions

- Speaker authentication and speaker recognition are different problems. Recognition is only a [weak] signal for “authenticating”.
- Speaker authentication can be broken if the attacker has speech data of the target and knows the authentication prompt.
- Although most TTS systems require 24 hours of speech to train, transfer learning is an effective way to reduce that time to an amount realistic for an attacker to abuse. Transfer learning is effective for reducing data requirements generally.
- In conclusion, it's relatively easy to spoof someone's voice
  - Will only get easier over time

# Final note

arXiv:1806.04558v1 [cs.CL] 12 Jun 2018

---

## Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis

---

**Ye Jia \***    **Yu Zhang \***    **Ron J. Weiss \***    **Quan Wang**    **Jonathan Shen**    **Fei Ren**

**Zhifeng Chen**    **Patrick Nguyen**    **Ruoming Pang**    **Ignacio Lopez Moreno**

**Yonghui Wu**

Google Inc.

{jiaye,ngyuzh,ronw}@google.com

*Be afraid. Be very afraid.*



**John Seymour and Azeem Aqil**