The State of Speech Recognition on Mobile


NCC-1701-A

The future won't be like Star Trek.

Scott Adams, creator of Dilbert

Why do I care about speech rec?


+


= Cape Bretoner

Here’s a conversation between two Cape Bretoners

P1: jeet?

P2: naw, jew?

P1: naw, t'rly t'eet bye.

And here’s the translation

P1: jeet?

P1: Did you eat?

P2: naw, jew?

P2: No, did you?

P1: naw, t'rly t'eet bye.

P1: No, it's too early to eat buddy.

Regular Alphabet
26 letters
Cape Breton
Alphabet
12 letters!

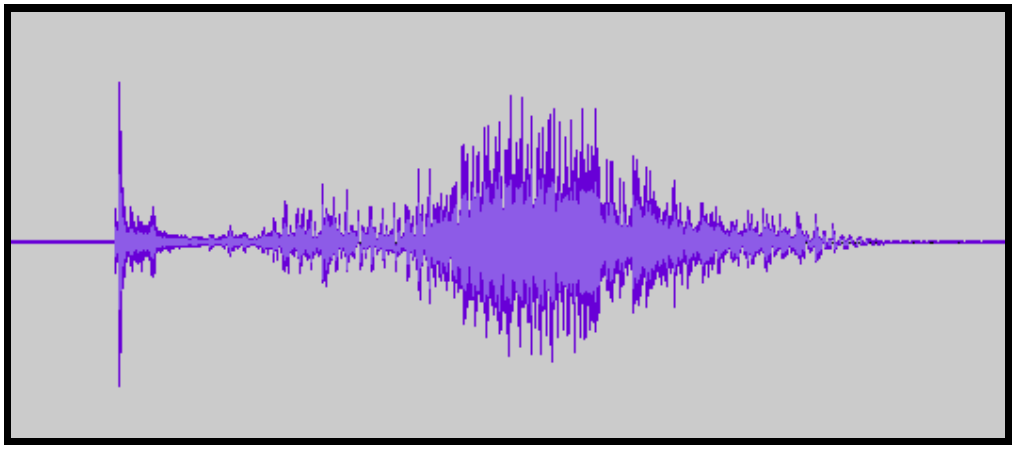
Alright, enough about me

What is speech recognition?

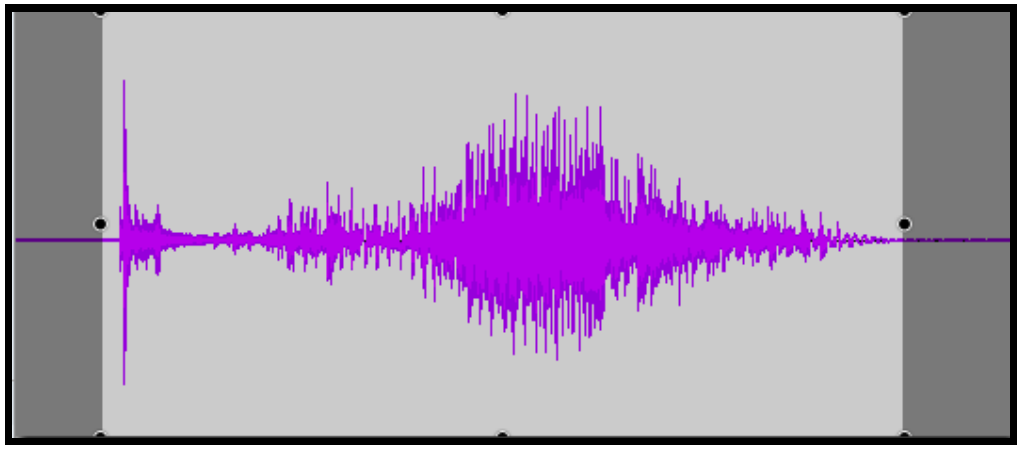
Speech recognition is the process of translating the spoken word into text.

The process of speech rec includes...

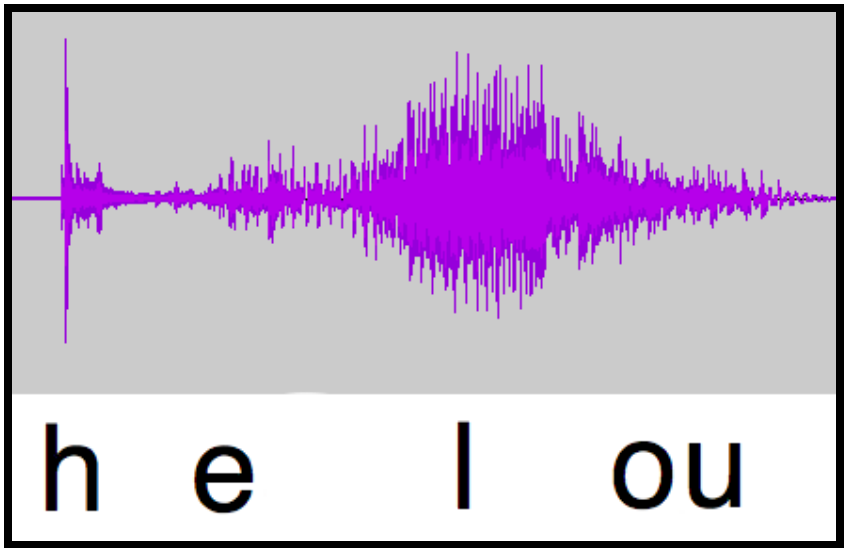
Record and digitize the audio data



Perform end pointing (trimming)



Split data into phonemes



What is a phoneme?

It is a perceptually distinct units of sound in a specified language that distinguish one word from another.

The English language has 44 distinct sounds

vowels

IPA	ASCII	examples
ʌ	^	cup, luck
ɑː	a:	arm, father
æ	@	cat, black
ə	..	away, cinema
e	e	met, bed
ɜːʳ	e:(r)	turn, learn
ɪ	i	hit, sitting
iː	i:	see, heat
ɒ	o	hot, rock
ɔː	o:	call, four
ʊ	u	put, could
uː	u:	blue, food
aɪ	ai	five, eye
aʊ	au	now, out
oʊ/əʊ	Ou	go, home
eəʳ	e..(r)	where, air
eɪ	ei	say, eight
ɪəʳ	i..(r)	near, here
ɔɪ	oi	boy, join
ʊəʳ	u..(r)	pure, tourist

consonants

IPA	ASCII	examples
b	b	bad, lab
d	d	did, lady
f	f	find, if
g	g	give, flag
h	h	how, hello
j	j	yes, yellow
k	k	cat, back
l	l	leg, little
m	m	man, lemon
n	n	no, ten
ŋ	N	sing, finger
p	p	pet, map
r	r	red, try
s	s	sun, miss
ʃ	S	she, crash
t	t	tea, getting
tʃ	tS	check, church
θ	th	think, both
ð	TH	this, mother
v	v	voice, five
w	w	wet, window
z	z	zoo, lazy
ʒ	Z	pleasure, vision
dʒ	dZ	just, large

Source: English language phoneme chart

By comparison, the Rotokas speakers in Papua New Guinea have 11 phonemes.

But the !Xóõ speakers who mostly live in Botswana have 112 phonemes.

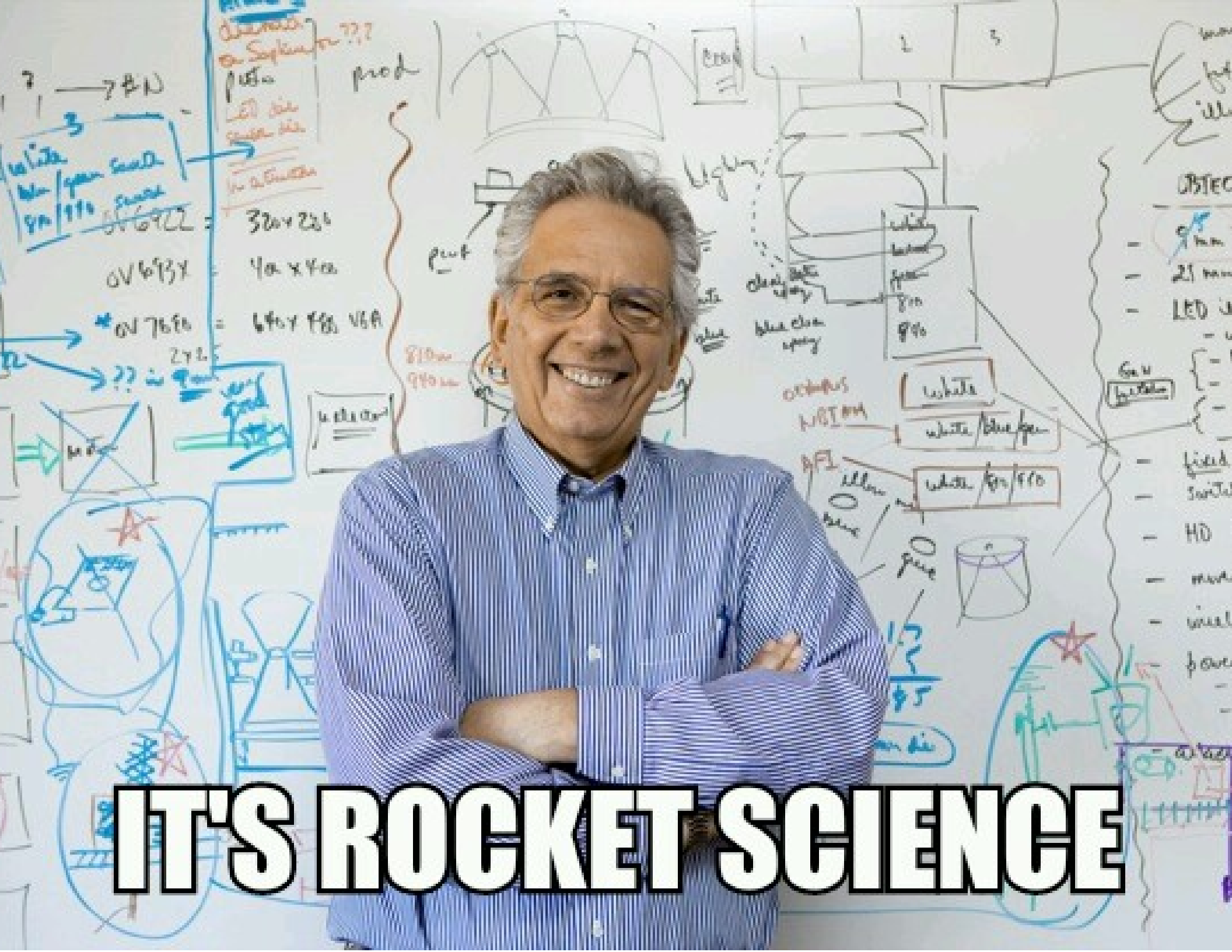
Apply the phonemes to the recognition model. This is a massive lexicon which takes into account all of the different ways words can be pronounced.

Analyze the results against the grammar

Return a confidence weighted result


```
[
  {
    "confidence": 0.97335243225098,
    "transcript": "hello"
  },
  {
    "confidence": 0.19940405040800,
    "transcript": "hell low"
  },
  {
    "confidence": 0.19910827091000,
    "transcript": "how low"
  }
]
```

Basically...


IT'S ROCKET SCIENCE

We want it to be like this

but more often than not...

Why is that?

When two people talk comprehension rates are better than 97%

A really good english language speech recognition system is right 92% of the time

Where does that extra 5% in error rate come from?

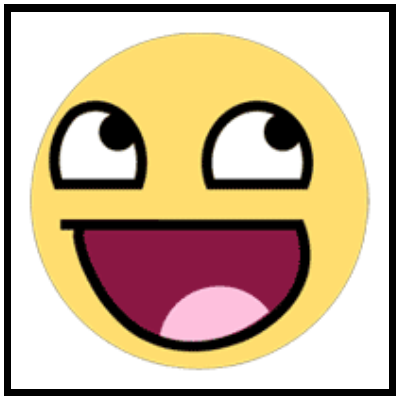
- Vocabulary size and confusability
- Speaker dependence vs independence
- Isolated or continuous speech
- Initiated vs spontaneous speech
- Adverse conditions

Mobile Speech Recognition

OS	Application	SDK
Android	Google Now	Java API
iOS	Siri	Many 3rd party Obj-C SDK's
Windows Phone	Cortana	C# API

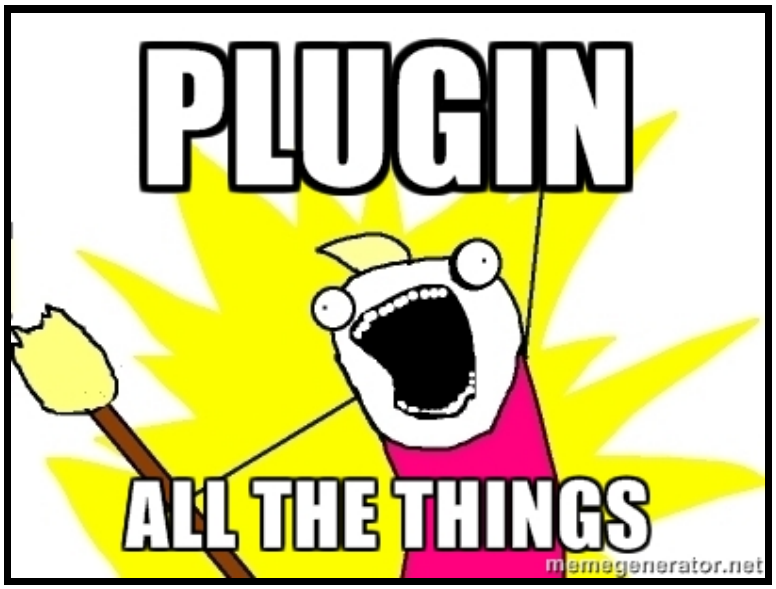
So how do we add speech rec to our app?

You may look at the W3C Speech API Specification



but only Chrome on the desktop has implemented that spec

But that's okay!



The spec looks like this:

```
interface SpeechRecognition : EventTarget {
    // recognition parameters
    attribute SpeechGrammarList grammars;
    attribute DOMString lang;
    attribute boolean continuous;
    attribute boolean interimResults;
    attribute unsigned long maxAlternatives;
    attribute DOMString serviceURI;

    // methods to drive the speech interaction
    void start();
    void stop();
    void abort();
};
```

With additional event methods to control behaviour:

```
attribute EventHandler onaudiostart;
attribute EventHandler onsoundstart;
attribute EventHandler onspeechstart;
attribute EventHandler onspeechend;
attribute EventHandler onsoundend;
attribute EventHandler onaudioend;
attribute EventHandler onresult;
attribute EventHandler onnomatch;
attribute EventHandler onerror;
attribute EventHandler onstart;
attribute EventHandler onend;
```

Let’s recognize some speech

```
var recognition = new SpeechRecognition();
recognition.onresult = function(event) {
  if (event.results.length > 0) {
    var test1 = document.getElementById("test1");
    test1.innerHTML = event.results[0][0].transcript;
  }
};
recognition.start();
```

Replace me...

So that's pretty cool...

...if taking dictation gets you going

But I want to do something more exciting with the result

Let's do something a little less trivial

```
recognition.onresult = function(event) {
    var result = event.results[0][0].transcript;
    var music = document.getElementById("music");
    switch(result) {
        case "jazz":
            music.src="jazz.mp3";
            music.play();
            break;
        case "rock":
            music.src="rock.mp3";
            music.play();
            break;
        case "stop":
        default:
            music.pause();
    }
};
```

Click to Speak

Which seems much cooler to me

Let’s ask the web a question

Click to Speak

Works pretty good...
...but ugly!

Let’s style our button with some CSS

```
<a class="speechinput">
    <img src="images/mic.png">
</a>
```

+

```
#speechinput input {
        cursor:pointer;
        margin:auto;
        margin:15px;
        color:transparent;
        background-color:transparent;
        border:5px;
        width:15px;
        -webkit-transform: scale(3.0, 3.0);
}
```

= 

And we'll add some color using

Pure-CSS-Speech-Bubbles by Nicholas Gallagher

Then pull it all together!

But wait, why am I using my eyes like a sucker?

We'll output the answer using SpeechSynthesis

The SpeechSynthesis spec looks like this:

```
interface SpeechSynthesis {
      readonly attribute boolean pending;
      readonly attribute boolean speaking;
      readonly attribute boolean paused;

      void speak(SpeechSynthesisUtterance utterance);
      void cancel();
      void pause();
      void resume();
      SpeechSynthesisVoiceList getVoices();
    };
```

The SpeechSynthesisUtterance spec looks like this:

```
interface SpeechSynthesisUtterance : EventTarget {
      attribute DOMString text;
      attribute DOMString lang;
      attribute DOMString voiceURI;
      attribute float volume;
      attribute float rate;
      attribute float pitch;
    };
```

With additional event methods to control behaviour:

```
attribute EventHandler onstart;
attribute EventHandler onend;
attribute EventHandler onerror;
attribute EventHandler onpause;
attribute EventHandler onresume;
attribute EventHandler onmark;
attribute EventHandler onboundary;
```

Plugin repo’s

- SpeechRecognitionPlugin - https://github.com/macdonst/SpeechRecognitionPlugin
- SpeechSynthesisPlugin - https://github.com/macdonst/SpeechSynthesisPlugin

Availability

OS	Recognition	Synthesis
Android	✓	✓
iOS*	Active development	Native to iOS 7.0
Windows Phone	×	×

* Working with Julio César (@jcesarmobile) to get iOS done

Getting started

```
cordova create speech com.example.speech speech
cd speech
cordova build android
cordova local plugin add https://github.com/macdonst/SpeechRecognitionPlugin
cordova local plugin add https://github.com/macdonst/SpeechSynthesisPlugin
cordova install android
```

For more information on hybrid applications

Check out Christophe Coenraets presentation on *Creating Native-Like Mobile Apps with AngularJS, Ionic and Cordova* 3:00pm today right here in Salon C.

But wait, one more thing...

Speech recognition and speech synthesis are not well supported in the emulator

and sometimes developing on the device can be a bit of a pain.

That's why I coded speechshim.js

https://github.com/macdonst/SpeechShim

Chrome + speechshim.js
=
W3C Web Speech API on your desktop