

Harnessing the power of AI in Text to Speech

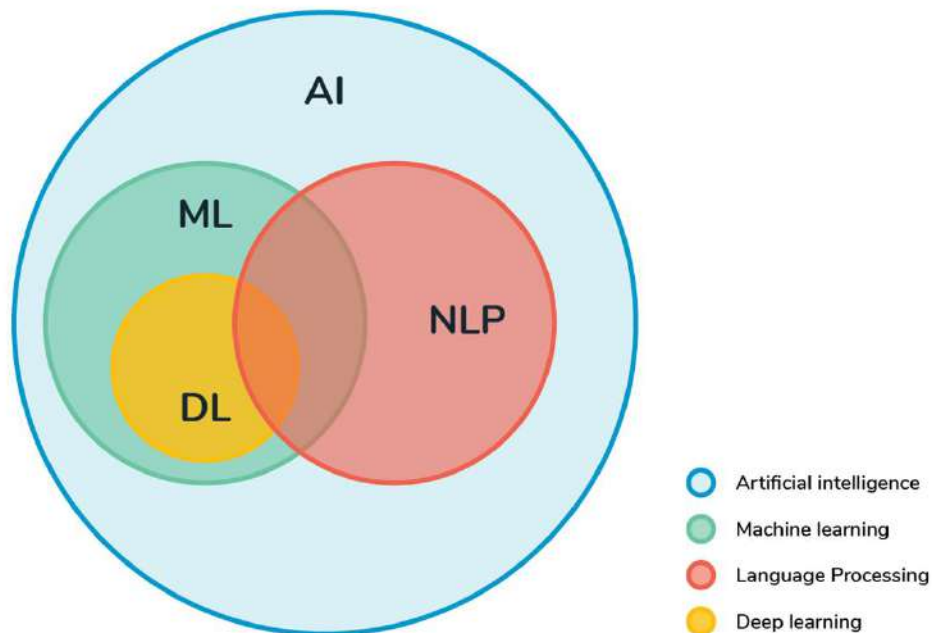
Using TTS in 2021 and Beyond

Overview

- What is Text To Speech [TTS]?
- Why is spoken language difficult?
- What to consider with TTS
- Let's make some TTS
- Providers

A word about AI

Artificial Intelligence is a large area of study, and primarily has seen Language Processing as a Data and Statistical problem. NLP is the branch that overlaps with what we see today in Machine Translation and Speech.



A historical progression

Every time I fire a linguist, the performance of our speech recognition system goes up.
-Fred Jelink 1985

1950

Starting out

Originally from the 30's Speech synthesis. It works but not always



1980

Statistics

Calculating statistical probabilities leads to better results.



2010

The age of the Neural Network

Neural Networks, TacoTron, WaveNet, Machine Learning and more Neural Networks



2020

Today

1-shot from hearing to speaking
Robust cloud models
Voice Agents everywhere

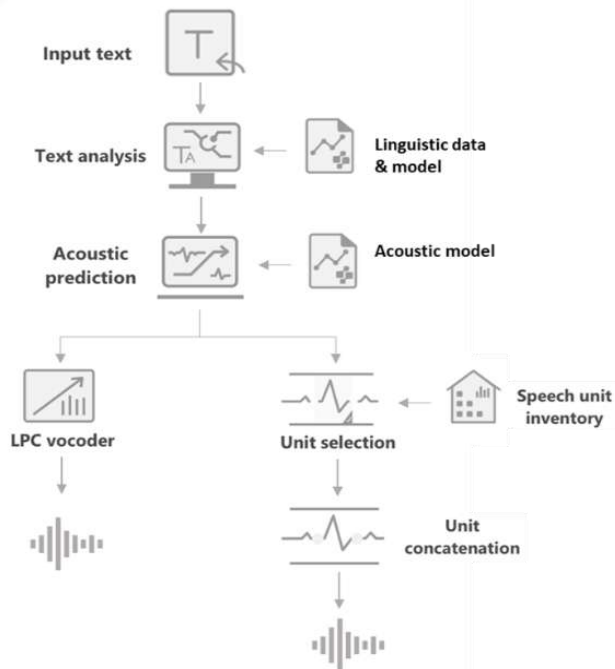
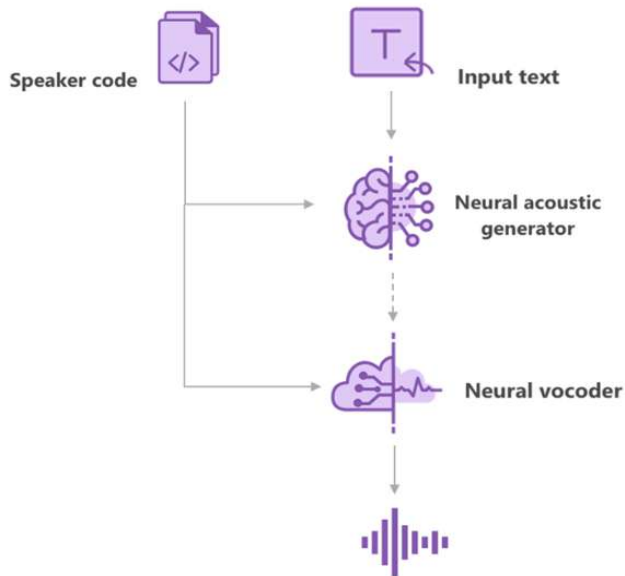


Neural TTS

Traditional TTS



- Joint optimization of pronunciation and prosody + high-fidelity audio generation
- Learning from large datasets across speakers



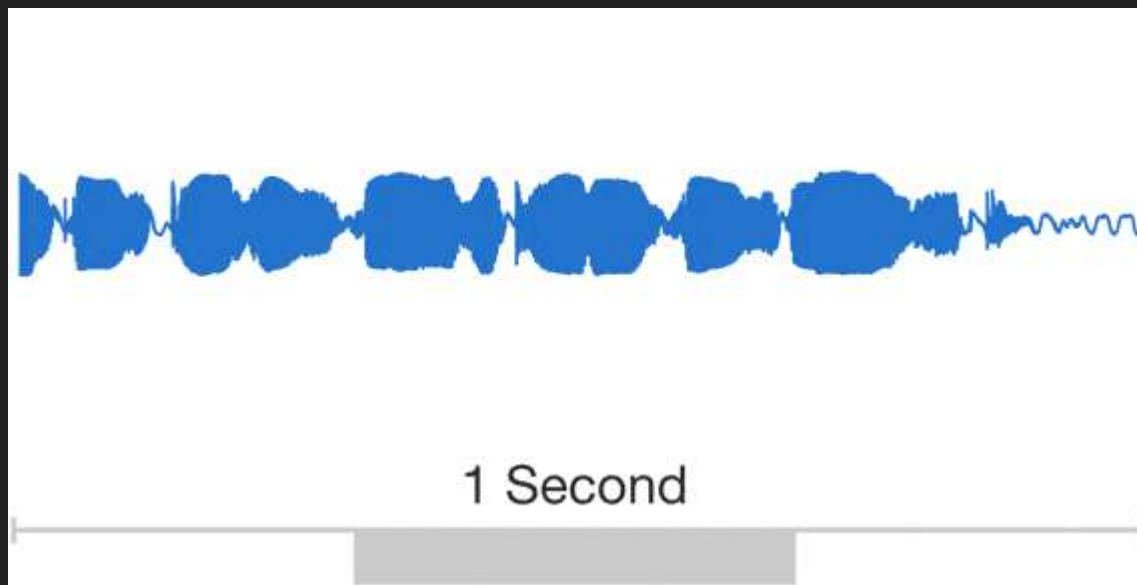
2008-2016: many changes in the AI space

We moved from procedural statistical processing to massively parallel processing.



Text To Speech is difficult

Speech is made up of sounds, and those sounds are complicated:



Voice is different...

1. watt ewe right iz knot watt u saye.
2. Even if it's written correctly you can't tell Bass from Bass...
3. The tools are theoretical
4. No existing "spell-check" or make-language-sound-right-check--human validation required
5. "Standards" vary by language/grammar/transcription method, and more!
6. Everyone has a slightly different implementation and acceptance criteria
7. Not all customers accept

Voice is faster and cheaper

- 1) Never gets sick
- 2) Sounds the same months later--Infinite retakes!
- 3) Doesn't get tired
- 4) Doesn't care what it is saying
- 5) Engineering costs + QA costs vs. Raw Talent

Prosody/Suprasegmentals [Phonetics 101]

Speech is more than just sounds put together, it's about *how* they are put together.

Suprasegmental: A set of qualities **super**imposed on a set of phonetic **segments**

- Pitch / Tone
- Juncture [e.g. punctuation]
- Stress [loud/soft]
- Intonation/Rhythm/Melody/pauses
- Duration

AKA: **Prosody**

"I never said that she stole my money"

Sounds versus Words

Localization has traditionally been text-based.

When we do voice--it's a human actor, reading a written script.

How many CAT tools exist vs. how many “language translation speech tools” exist.

Text has many advantages, until recently we didn't have similar methods for spoken language and our tools have been limited.

Modern speech tools are a product of the AI/NLP Evolution of the past 5 years

IPA and SSML

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2018)

CONSONANTS (PULMONIC)

© 2018 IPA

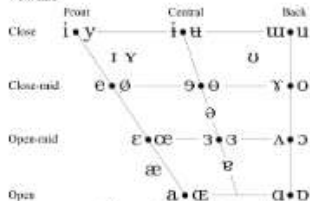
	Bilabial	Labiodental	Dental	Alveolar/Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d	ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n	ɳ	ɲ	ŋ	ɴ		
Fricill				r				ʀ		
Tap or Flap				ɾ		ɽ				
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ						
Approximant		ʋ		ɹ		ɻ	ɰ			
Lateral approximant				l		ɭ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

	Clicks	Voiced implosives	Ejectives
ʘ	Bilabial	ɓ	ɰ
ǀ	Dental	ɗ	ɰ
ǃ	Postalveolar	ɟ	ɰ
ǂ	Palatoalveolar	ɠ	ɰ
ǁ	Alveolar lateral	ɠ	ɰ

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

- ʌ Vowelless labial-velar fricative
- ɕ ʑ Alveolo-palatal fricative
- ʋ Vowelless labial-velar approximant
- ɹ Vowelless alveolar lateral fricative
- ɰ Vowelless labial-palatal approximant
- ɰ Vowelless labial-palatal approximant
- ɰ Vowelless epiglottal fricative
- ɰ Vowelless epiglottal fricative
- ʕ Epiglottal plosive

ɬ ɮ

DIACRITICS Some diacritics may be placed above a symbol with a descender, e.g. ɰ̥

◌̥	Voiceless	◌̄	Acute accent	◌̂	Open-mid central	◌̃	Open-mid central	◌̆	Open-mid central
◌̇	Voiceless	◌̈	Open-mid central	◌̉	Open-mid central	◌̊	Open-mid central	◌̋	Open-mid central
◌̌	Aspirated	◌̍	Aspirated	◌̎	Aspirated	◌̏	Aspirated	◌̐	Aspirated
◌̑	More rounded	◌̒	More rounded	◌̓	More rounded	◌̔	More rounded	◌̕	More rounded
◌̖	Low rounded	◌̗	Low rounded	◌̘	Low rounded	◌̙	Low rounded	◌̚	Low rounded
◌̜	Advanced	◌̝	Advanced	◌̞	Advanced	◌̟	Advanced	◌̠	Advanced
◌̢	Retracted	◌̣	Retracted	◌̤	Retracted	◌̥	Retracted	◌̦	Retracted
◌̧	Compressed	◌̨	Compressed	◌̩	Compressed	◌̪	Compressed	◌̫	Compressed
◌̬	Mid-centralized	◌̭	Mid-centralized	◌̮	Mid-centralized	◌̯	Mid-centralized	◌̰	Mid-centralized
◌̱	Syllabic	◌̲	Syllabic	◌̳	Syllabic	◌̴	Syllabic	◌̵	Syllabic
◌̶	Non-syllabic	◌̷	Non-syllabic	◌̸	Non-syllabic	◌̹	Non-syllabic	◌̺	Non-syllabic
◌̻	Rhoticity	◌̼	Rhoticity	◌̽	Rhoticity	◌̾	Rhoticity	◌̿	Rhoticity

1886

India pale ale



Beer style



India pale ale is a hoppy beer style within the broader category of pale ale. The export style of pale ale, which had become known as India pale ale, developed in England around 1840, later became a popular product there. [Wikipedia](#)

Alcohol by volume: 4.5% - 12.1%

Original Gravity: 1.050 - 1.075

Bitterness (IBU): 40 - 120

Color (SRM): 6 - 14

Final Gravity: 1.010 - 1.018

1840

SSML

It's like HTML, but for speech!

Allows you to control timing, volume, pitch, stress... Prosody

Each vendor has a slightly different implementation but there is an official standard from the W3 consortium.

2010

<https://www.w3.org/TR/speech-synthesis11/>

Voxabot TTS Editor

Languages ▾ Voices ▾

Arabic, Egypt

zh-CN-HuihuiRUS

zh-CN-Kangkang-Apollo

zh-CN-XiaoxiaoNeural ✓

Custom

Tools

SSML

Custom tag

Bre

Emphasis >

Language

Pitch >

Rate >

Volume >

Phoneme

Say-as >

Alias

Sentence

Custom tag



Enter tag name

Enter attribute name

Enter attribute value

Cancel

Ok

Custom SSML
Tag

度。 </p>



SSML Applied

Edit SSML Custom Tools Connections

← → I* ⏏ Languages ▾ Voices ▾

坡下立着一只鹅，坡下就是一条河。

坡下立着一只鹅，坡下就是一条河。

宽宽的河， (moderate) 肥肥的鹅， (medium) 鹅要过河， (medium) 河要渡鹅， 不知是鹅过河， 还是河渡鹅。

```
< speak version="1.0" xml:lang="zh-CN" >
< voice name="zh-CN-XiaoxiaoNeural" >
< p >坡下立着一只鹅，坡下就是一条河。 </p >
< p >宽宽的河， {emphasis level="moderate"}肥肥的鹅</emphasis>， {prosody rate="medium"}鹅要过河， {prosody pitch="medium"}河要渡鹅</prosody>， {prosody rate="medium"}不知是鹅过河</prosody>， {prosody rate="medium"}还是河渡鹅。 </p >
</ voice >
</ speak >
```

Voxabot Demo

Working with TTS in many languages

TTS is synthesized and generated based on abstract language rules--abstract because sometimes it's a guess.

Adding tags and punctuation can make unknown changes--and also fix defects.

Since most people don't speak 120+ languages, you send it to a linguist who can tell you where it's wrong.

TTS feedback: How To

Just like any feedback:

- 1) Be precise with replace [this] with [that]
- 2) If the error is phonetic as in Bass vs. Bass [fish/instrument] then spell it out like:
 - a) BASSE or BASE
 - b) If you have pinyin or other phonetic guides use them:

```
<say>  
你说 <phoneme alphabet="x-amazon-pinyin" ph="bo2">薄</phoneme>。  
我说 <phoneme alphabet="x-amazon-pinyin" ph="bao2">薄</phoneme>。  
</say>
```

Main TTS providers

Amazon: AWS Polly

Microsoft: Azure Cognitive Speech Services

Google: Google Compute Cloud: Speech

Each provides a set of languages and are gradually expanding their Neural TTS offerings.

Other players exist in Canada, US and Australia which are building their own models.

The Future

Voice cloning [actors go virtual]

Customized voices [brand voice]

No Translation stage [voice to voice]

Conclusion

TTS is easy, language is hard

You can almost replace voiceover.