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

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

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

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

Traditional TTS

- Input text
- Text analysis
- Acoustic prediction
- Acoustic model
- LPC vocoder
- Unit selection
- Unit concatenation

Linguistic data & model

Speech unit inventory

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:

https://www.deepmind.com/blog/article/wavenet-generative-model-raw-audio
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 superimposed 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
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
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.

https://www.w3.org/TR/speech-synthesis11/
Voxabot TTS Editor
SSML Applied
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:

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