HELLO!

VOYSIS

NOW WE'RE TALKING

Text-to-Speech: Technologies of Tomorrow

Presented By: Dr. Peter Cahill
April 2015

Motivations

- Text-to-Speech has been around for decades, why is it not solved yet?
- Has it matured or is it still developing?
- What are the next developments going to be?
- Traditionally, technology limitations have been the barrier for text-to-speech to be used in more markets.

Text-to-Speech

Machines (or devices) speak to people.

Ideally machines could speak expressively, clearly, and consistently, in any language, dialect, accent or voice.

Reading text is a creative problem.
Professionals who do this are actors.

“I turned and saw that Flora, whom, ten minutes before, I had established in the schoolroom with a sheet of white paper, a pencil, and a copy of nice "round o's," now presented herself to view at the open door. She expressed in her little way an extraordinary detachment from disagreeable duties, looking at me, however, with a great childish light that seemed to offer it as a mere result of the affection she had conceived for my person, which had rendered necessary that she should follow me. I needed nothing more than this to feel the full force of Mrs. Grose's comparison, and, catching my pupil in my arms, covered her with kisses in which there was a sob of atonement.”

State-of-the-art

Data Requirements

- SOTA systems require many carefully crafted datasets:
  - 1 dataset per voice (larger the better, common to have ~100 hours for SOTA). Ideally in domain.
  - Pronunciation data (how to pronounce 100,000's of words in a language)
  - Grammatical or syntax data (for understanding text)
  - + more!

Technologies

- Increasingly dependent on machine learning
- Still very sensitive to errors in datasets
- Significant manual tuning required per voice to achieve optimal results
- Speed is a problem (computers not fast enough!)

State of the art

- Highly tuned purpose built systems actually sound quite good.
- Adding languages or voices is expensive (time + cash)

Recent/future developments

NOW WE'RE TALKING

Foreign Pronunciations

- Many names have a foreign origin.
- It can sound terrible if the system tries to pronounce a foreign word as if it were English.
- How would you pronounce these names (Irish origin): Aoife, Niamh, Caoimhe, Sinead?
- Pronouncing a word with it's correct foreign pronunciation may not be the best thing to do.
- Another option is to pronounce it as a speaker of the current language would.

Vocoders (Control)

- Decompose a speech signal into specific parts. These parts can be modified and then new speech can be generated.
- Traditionally we’ve had a choice when it comes to speech quality: natural sounding (but error prone) or robotic (but reliable).
- Once we modify natural speech, it sounds more robotic.
  - This is a challenge for systems that need to sound natural. Sounding natural some of the time is generally not acceptable.
- Recent technologies (TANDEM-STRAIGHT, Vocaine) enable more modifications with less of a robotic sound.

Expressive Voices

- Current approaches use multiple voice datasets, “happy”, “sad”, etc.
- Either switch between which dataset is being used, or model the expressions and blend.
- Multiple challenges:
  - How to obtain data?
  - How to classify expressions? (sad vs bored)
  - When to sound expressive? (using the wrong expression may be more damaging than sounding neutral).

VOYSIS

Unsupervised

- There's a lot of voice data out there (TV channels, radio stations, online video, audio books, phone calls, etc.)
- Can we build a machine that can learn every language, voice, dialect, accent **by itself**?
  - Can only use existing (unstructured) audio data or text.

Unsupervised

- Relatively new approach
- Many traditional components are no longer useful
- If we can model these data sources, maybe then we can capture and model truly expressive speech
- and support 100s of languages…
- …and 1000s of voices

Conclusion

These new technologies will open up new markets.

There has been a lot of technology developments in the past ~3 years. It looks like this trend will continue.

The TTS we imagine in the future may not be so far away.

THANKS!

peter@voysis.com

VOYSIS

NOW WE'RE TALKING

VOYSIS.COM

@voysis.com