Spontaneous conversational speech synthesis

The making of a podcast voice – breathing, uhs & ums and some ponderings about appropriateness

Éva Székely, Gustav Eje Henter, Jonas Beskow, Joakim Gustafson
Why synthesise spontaneous conversational speech?
# Types of TTS corpora

<table>
<thead>
<tr>
<th></th>
<th>Traditional TTS corpus</th>
<th>Public domain Audiobook</th>
<th>Public domain Conversational Podcast</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recording conditions</strong></td>
<td>Controlled</td>
<td>Uncontrolled</td>
<td>Uncontrolled</td>
</tr>
<tr>
<td><strong>Type of speech</strong></td>
<td>Read / acted</td>
<td>Read / acted</td>
<td>Spontaneous / semi-planned</td>
</tr>
<tr>
<td><strong>Amount of data</strong></td>
<td>Limited to resources</td>
<td>Unlimited</td>
<td>Unlimited</td>
</tr>
<tr>
<td><strong>Transcriptions</strong></td>
<td>Available</td>
<td>Available</td>
<td>Not available</td>
</tr>
<tr>
<td><strong>Segmentation to utterances</strong></td>
<td>Decided on pre-recording</td>
<td>Sentence / paragraph level post-recording</td>
<td>Not existent</td>
</tr>
<tr>
<td><strong>Nr of speakers</strong></td>
<td>1</td>
<td>1</td>
<td>2 or more</td>
</tr>
<tr>
<td><strong>Disfluencies</strong></td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>
Detecting breath events and overlapping speech for segmentation & utterance selection
CNN-LSTM speaker dependent breath detector
ThinkComputers Corpus (TCC)

- Weekly podcast
- Tech news, product reviews
- 150 h available copyright free
- No transcriptions
- 2 speakers mixed into a single channel

Speaker with most air time was selected
27 episodes -> 9 hours of clean breath groups
Transcription

Fully automatic using ASR and forced alignment:

**Google Speech API**, video model – best accuracy

**IBM Watson Speech to Text** – generic hesitation label

**Gentle forced aligner** – distinguish between *uh* & *um*

6218 breath groups orthographically transcribed, filled pauses *uh* and *um* identified and transcribed, other disfluencies indicated with a nr per utterance
Text-to-Speech

Tacotron 2 spectrogram prediction framework
waveform synthesis: Griffin-Lim algorithm

Pronunciation accuracy was improved by phone-level transcription and transfer learning with a read-speech voice


Perceptual evaluation: *appropriateness*

Goal: to see how the podcast voice was perceived, compared with voices trained on:

a) read speech

b) lab-recorded, manually annotated spontaneous conversational speech

“How well does the speaking style match the content of the utterance?”
Evaluation of appropriateness for different genres

• Read speech: 10 utterances originating from the Arctic Corpus
• Public speaking: 10 utterances transcribed from political speeches and keynote talks; 7 of the prompts contain FPs
• Casual conversation: 20 utterances from a corpus of casual spontaneous conversations from a TTS corpus, 11 of the prompts contain FPs
Synthetic voices

- **POD-FP:** TC corpus; FPs transcribed
- **POD-FLU:** Fluent breath groups from TC; no FPs
- **READ-PR:** Voice trained on audiobook data; no FPs
- **GAME-FP:** Lab-recorded spontaneous speech; FPs transcribed
Mushra-like listening test results

“How well does the speaking style match the content of the utterance?”

Samples at: http://www.speech.kth.se/tts-demos
Evaluations from a perceptual point of view

Can we use speech synthesis to understand perceptual aspects of spontaneous speech phenomena?
Preference tests on the impact of filled pauses

2 Interspeech keynotes 10-27 second long paragraphs with and without FPs:

Which speaker sounds more **engaging**?

POD-PT-FP  no diff.  POD-PT-FLU

Which speaker sounds more **authentic**?

POD-PT-FP* (p=0.007)  no diff.  POD-PT-FLU
“Would you want a robot to sound hesitant? Why or why not?”

<table>
<thead>
<tr>
<th>Yes – 45%</th>
<th>Undecided – 19%</th>
<th>No – 36%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes, sounds more authentic and genuine.</td>
<td>indifferent</td>
<td>No, because it would sound more human-like.</td>
</tr>
<tr>
<td>sounds far more realistic</td>
<td>Unsure, closer to no because I don't think it's really necessary but it also depends on the reason for it.</td>
<td>not really</td>
</tr>
<tr>
<td>Yes, more realistic</td>
<td>I have no idea.</td>
<td>No, it might make it too human and be somewhat uncanny valley</td>
</tr>
<tr>
<td>Yes, it makes it more human and relatable.</td>
<td>As long as I knew it was a robot and was not done without my knowledge</td>
<td>No. It is not a good idea I think.</td>
</tr>
<tr>
<td>Yes so it sounds more like a person and more relatable.</td>
<td>Only if it was trying to fake being a human. Otherwise it sounds too artificial.</td>
<td>No, because it would sound too human like. Over the phone, I wouldn't be able to tell I am talking to a robot.</td>
</tr>
<tr>
<td>Yes, gives human factor</td>
<td>it would not bother me to be honest.</td>
<td>No - Robots should know exact answers</td>
</tr>
<tr>
<td>yes as its more realistic as a humans voice</td>
<td>Yes, and no. Yes, because it makes it sound more like a regular person but ultimately no, because it can be harder to listen to and transcribe.</td>
<td>No, it would possibly make it more difficult to determine whether you're talking to real person or a machine.</td>
</tr>
<tr>
<td>Yes, it does sound a bit more human that way.</td>
<td>if it makes it sound more real</td>
<td>No, I feel uncomfortable blurring the lines between what sounds naturally human and what is machine.</td>
</tr>
<tr>
<td>Yes, makes it seem more human.</td>
<td></td>
<td>No. I want robots to sound authoritative.</td>
</tr>
<tr>
<td>Yes more real effect</td>
<td></td>
<td>No as I'm not sure it's an important factor</td>
</tr>
<tr>
<td>Yes, because it would sound more human.</td>
<td></td>
<td>No as I would want it to speak correctly at all times.</td>
</tr>
</tbody>
</table>
"Would you want a robot to sound hesitant? Why or why not?"

<table>
<thead>
<tr>
<th>Yes – 45%</th>
<th>Undecided – 19%</th>
<th>No – 36%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes as it sounds more like a human voice</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>yes, its more realistic and makes me feel like i'm listening to a real person.</td>
<td></td>
<td>No so you know its a robot</td>
</tr>
<tr>
<td>Yes if the goal was to make it sound human. It all depends upon the use of the robot. Customer service robot then yes but maybe an automated robotic vacuum then maybe no. Context would be key.</td>
<td></td>
<td>No. I like the audio clear cut between machine and human. Perhaps though it is important in other cultures or languages to have different intonation and beats in order to get information across efficiently.</td>
</tr>
<tr>
<td>Yes, much more easy to listen to for prolonged periods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I think it's comforting to have a hesitant voice from them - things like phone calls make me anxious and having something more human sounding on the other end makes it more comfortable.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes sounds more human</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes as its more realistic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Why not? More human-like sounds like a very good idea</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Take-home messages:

1. We can automatically process and annotate found speech data accurately enough for TTS.

2. Realistic TTS is more genre-dependent than citation style synthesis → new research directions.

3. More control is needed, but we can already conduct experiments that reveal perceptual aspects of spontaneous speech phenomena such as disfluencies, which would have been difficult using natural speech.
More details in:


Questions?
Interchangeability of *uh* and *um*

Prompts from recorded discussions on the design of a remote control (AMI corpus)

3 conditions:
1) FP type copied from original recording
2) FP opposite type as original
3) FP type decided automatically, by merging the two labels into one before training the voice
Perceptual test

• Pairwise evaluation between the three conditions.
• Listeners were made aware they were evaluating synthetic speech with FPs and asked which one hesitated more realistically.
• They also had the option to select that both are plausible or neither is plausible.
Results

• 69% of all utterances were considered realistic, and only in 7% of the cases was neither considered realistic.

• Overall, there was no significant difference between copying the type of FP or using the opposite.

• Letting the system decide outperformed the other two options, in particular for FPs in the middle of the utterance.
Demo
Off the cuff: Extemporaneous speech delivery with TTS

Extemporaneous speech: a type of public speaking which uses a structured outline but is otherwise delivered conversationally, off the cuff.

TTS evaluation from the production point of view: *What if you are the speaker, not the listener?*

Is it possible to simulate responsiveness to audience with spontaneous TTS?
Evaluation of pronunciation accuracy

Improved by phone-level transcription and transfer learning with a read-speech voice

<table>
<thead>
<tr>
<th>Voice</th>
<th>Nr of pronunciation errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grapheme-level input with transfer learning</td>
<td>49</td>
</tr>
<tr>
<td>Phoneme-level input and random initialisation</td>
<td>43</td>
</tr>
<tr>
<td>Phoneme-level input with transfer learning</td>
<td>13</td>
</tr>
</tbody>
</table>

Pronunciation assessment of 400 Harvard sentences.
Perceptual evaluation of fluent speech
Next steps

Breath as an *input* feature for implicit prosody control

- Traditional TTS corpora: utterance length is given, the reader adjusts breathing
- Spontaneous BG corpus: dynamic relationship between breath and length of utterance, related to speech planning
Next steps

Further evaluating TTS in the context of *appropriateness*:

How does speech transfer across genres?

Citation-style TTS did a mediocre job on everything, but now genre transfer starts to matter.
um

original  synthesis

uh

original  synthesis
Length of uh and um in different BG positions

**original (held-out)**

- **start**
- **middle**
- **end**

**synthesis**

- **start**
- **middle**
- **end**
Perceptual evaluation of disfluent speech
Conclusions

✓ Breath detection works well in extracting utterances out of messy dialog data for TTS corpus

✓ Spontaneous TTS beats TTS from read speech for spontaneous speech genres in terms of appropriateness

✓ We can relinquish control over FPs without quality loss
Overview

1. From podcast to TTS corpus: Breath detection
2. Spontaneous conversational speech synthesis
   • TTS and evaluation
   • How to deal with uhs and ums?
   • Demo

3. Future work
Filled pauses in spontaneous TTS

Previously:

A) Where should we put them?
B) How should they sound?
  • Conclusion: It is not enough to put them in the right place, they should also sound right.

Now:

C) What should we do with them?
  • Conclusion: As long as they sound right, it matters less where they are.
Third-Party Modified Version of the webMUSHRA Software

Hesitation by an AI voice

Which one hesitates more like a human does?

Stop

A  Play  B

Which version hesitates more realistically (like a human), or are they both equally plausible?

A  B  Both are plausible  Neither is plausible

Previous  Next
How to train your fillers?

1) How to gain control over uh and um when synthesising disfluent speech?

2) What is the best way to synthesise fluent speech out of a disfluent corpus?