SPONTANEOUS CONVERSATIONAL SPEECH SYNTHESIS FROM FOUND DATA

Éva Székely, Gustav Eje Henter, Jonas Beskow, Joakim Gustafson

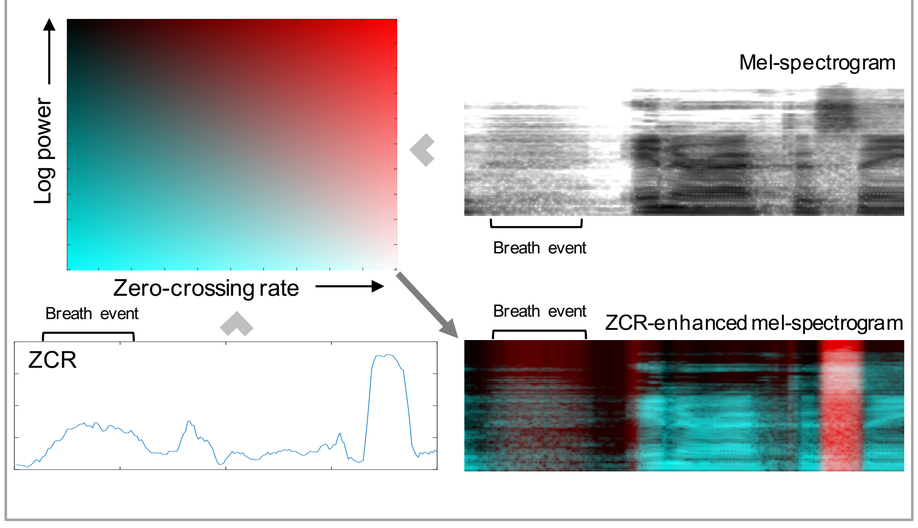
Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

Motivation: Most speech is spontaneous and conversational. TTS should be, too!

Approach

Data source: Public domain conversational podcast with 2 speakers
Issue 1: No punctuation/sentences → Segment on breath groups (BGs)
Issue 2: Multiple speakers → Use lightly-supervised speaker-dependent BG extractor as described in [1]
Issue 3: No transcription → Use ASR: Google Speech API, IBM Watson
Issue 4: Disfluent; filled pauses (FPs) → use Gentle forced aligner to transcribe FPs as 'uh' and 'um'
Result: 27 episodes → 9 hours of clean, single-speaker BGs
TTS: Rayhane Mama's Tacotron 2 [2] + Griffin-Lim

Segmentation of 1-hour long episodes into single-speaker utterances: with a CNN-LSTM breath detector on ZCR-enhanced spectrograms [1]



Evaluation 1: Formal evaluation on the effect of transfer learning and G2P front end on pronunciation

G2P	Transfer learning	Pron. errors
	✓	49
✓		43
✓	✓	**13***

Finding 1: Both transfer learning with a read-speech corpus, and G2P improve pronunciation.
2 evaluators on 400 phonetically balanced (Harvard) sentences

Listen here! →

www.speech.kth.se/tts-demos

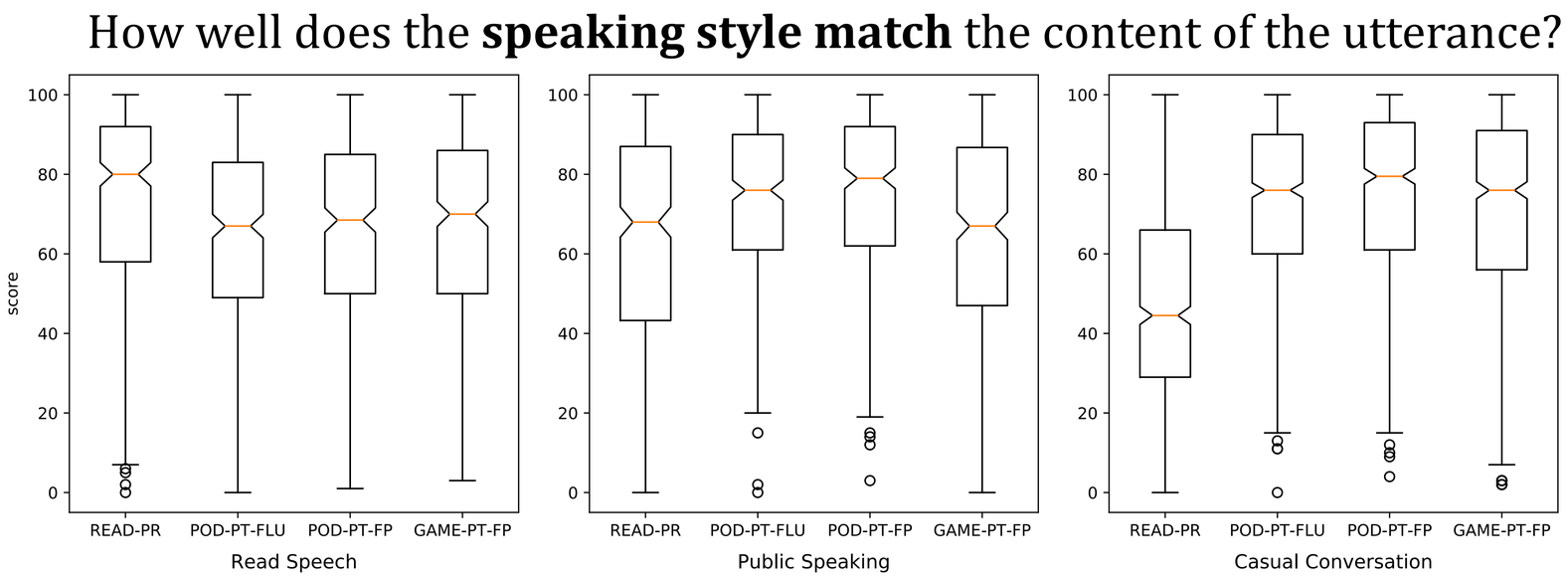
Evaluation 2: Disfluencies and conversational characteristics

Voice	Deletions	Repeated Syllables	*'the' as 'thee'*	*'a' as 'ei'*
Whole corpus	15	32	40/422	18/98
Fluent half	5	6	57/422	13/98

Finding 2: The disfluent voice has a significantly slower speech rate, more repeated syllables, but the distribution of non-reduced forms of determiners remains similar

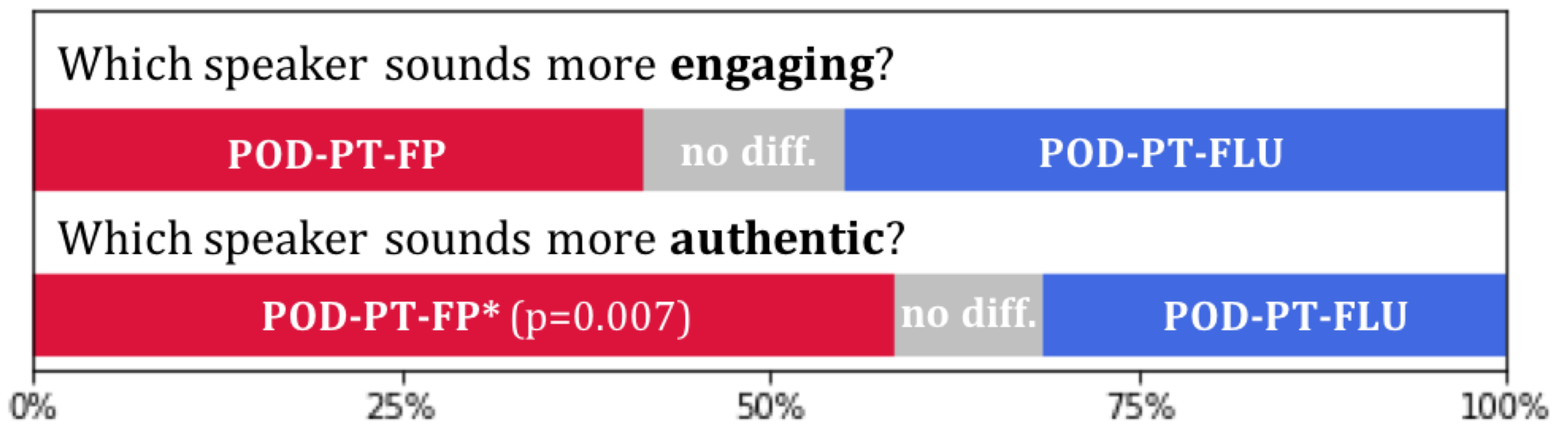
System	Data	Found	Spont.	Disfl.	Has text
READ-PR	LJ Speech	✓			✓
POD-PT-FLU	Podcast [1]	✓	✓		
POD-PT-FP	Podcast [1]	✓	✓	✓	
GAME-FT-FP	Conversation		✓	✓	✓

Evaluation 3: MUSHRA-like listening test on genre-appropriateness



Finding 3: On spontaneous text prompts, our found and spontaneous podcast TTS speaking style was preferred over TTS from either found but read, or spontaneous conversational but lab-recorded corpora.
No significant difference was found between the disfluent and the fluent podcast voices.

Evaluation 4: Preference test on the impact of filled pauses on perception



Finding 4: Having FPs ("FP" as opposed to "FLU") improved perceived authenticity when synthesising series of prompts from public speeches.
The presence of FPs seemed to have no effect on how engaging the listeners found the speaker.

This research was supported by the Swedish Research Council Project Incremental Text-To-Speech Conversion VR (2013-4935) and by the Swedish Foundation for Strategic Research project EACare (RIT15-0107). The authors would also like to thank the creators of the ThinkComputers podcast for making their recordings available in the public domain.

[1] É. Székely, G. E. Henter, and J. Gustafson, "Casting to corpus: Segmenting and selecting spontaneous dialogue for TTS with a CNN-LSTM speaker-dependent breath detector," in *Proc. ICASSP*, 2019, pp. 6925–6929.
[2] R. Mama, "Tacotron-2 Tensorflow implementation," https://github.com/Rayhane-mamah/Tacotron-2/, 2018.