SPONTANEOUS CONVERSATIONAL TTS 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 single-speaker BG extractor as described in [1]
Issue 3: No transcription → Use off-the-shelf ASR
Issue 4: Disfluent; filled pauses (FPs) → Annotate using additional tools
Result: 27 episodes → 9 hours of clean, single-speaker BGs
TTS: Rayhane Mama's Tacotron 2 [2] + Griffin-Lim

Listen here! →

www.speech.kth.se/tts-demos

Tool	WER	Um	Uh	Other disfl.
Google Cloud	Low	Omitted	Omitted	Omitted
IBM Watson	Slightly higher	"Hesitation"	"Hesitation"	Omitted
Gentle	- (Forced align)	"Um"	"Uh"	"Uh"

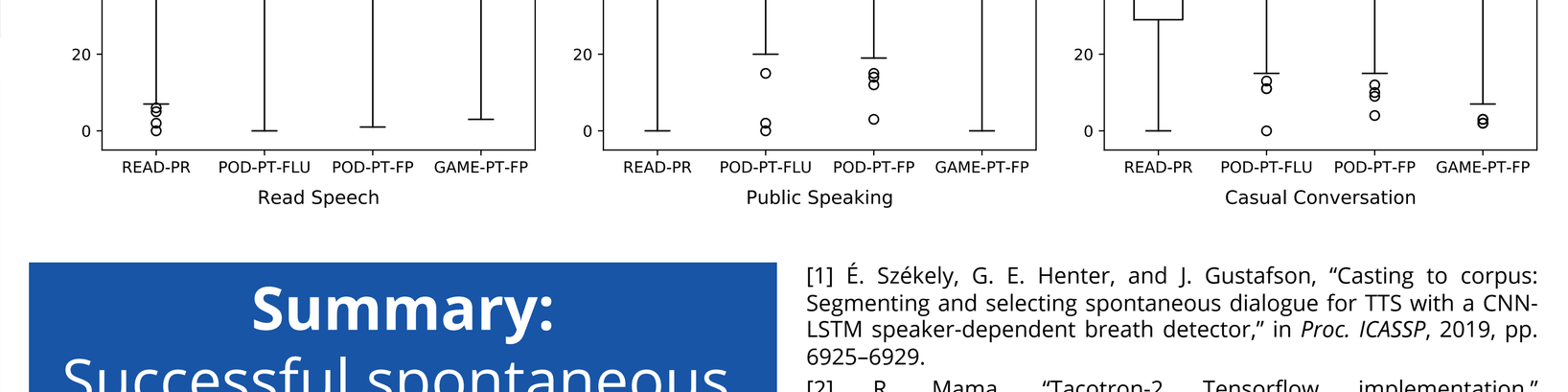
Finding 1: Transfer learning and G2P improve TTS pronunciation:

G2P	Transf. learn.	Pron. errs.
	✓	49
✓		43
✓	✓	**13***

Finding 2: Training with or without annotated FPs gives FP insertion with or without control:
Spontaneously disfluent TTS!

FP at B	M	E	Held out	TTS	*p*-val.
			49%	66%	<0.001
✓			23%	20%	0.109
	✓		17%	6%	<0.001
		✓	3%	5%	0.055
✓	✓		6%	1%	<0.001
✓		✓	1%	1%	0.844
	✓	✓	1%	1%	0.592
✓	✓	✓	0%	0%	0.200

Finding 3: Having FPs ("FP" as opposed to "FLU") improved perceived authenticity when synthesising series of public-speaking prompts:



Finding 4: On spontaneous text prompts, our found and spontaneous podcast TTS speaking style was preferred over TTS from either found but read, or spontaneous but lab-recorded, speech:
("POD" vs. "READ" = LJ Speech [3] and "GAME" [4], respectively)

Finding 5: Omitting disfluent utterances from training ("HalfFluent") or from fine-tuning ("TransFluent") gives more naturally fluent TTS:

Summary:
Successful spontaneous, conversational, found-data TTS with Tacotron 2

[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.

[3] K. Ito, "The LJ Speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[4] É. Székely, J. Mendelson, and J. Gustafson, "Synthesising uncertainty: the interplay of vocal effort and hesitation disfluencies," *Proc. Interspeech*, 2017, pp. 804–808, 2017.

This poster is based on: É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Spontaneous conversational speech synthesis from found data," to appear at *Interspeech 2019*; and É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," submitted to *SSW10*, 2019.

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.