HOW TO TRAIN YOUR FILLERS: UH AND UM IN SPONTANEOUS SPEECH SYNTHESIS

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

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

Using spontaneous conversational speech for TTS raises questions on how disfluencies such as filled pauses (FPs) should be approached. Detailed annotation of FPs in training data enables precise control at synthesis time; coarse or non-existent FP annotation, when combined with stochastic attention-based neural TTS, leads to synthesisers that insert these phenomena into fluent prompts on their own accord.
In this study we investigate, objectively and subjectively, the effects of FP annotation and the impact of relinquishing control over FPs in a Tacotron TTS system built from a conversational podcast corpus.

ThinkComputers Corpus (TCC)
- Weekly tech podcast, spontaneous conversational speech
- Male AE speaker
- 9h, segmented to single-speaker breath groups [1]
- Automatic transcription with ASR and Gentle forced aligner

TTS: Tacotron + Griffin-Lim

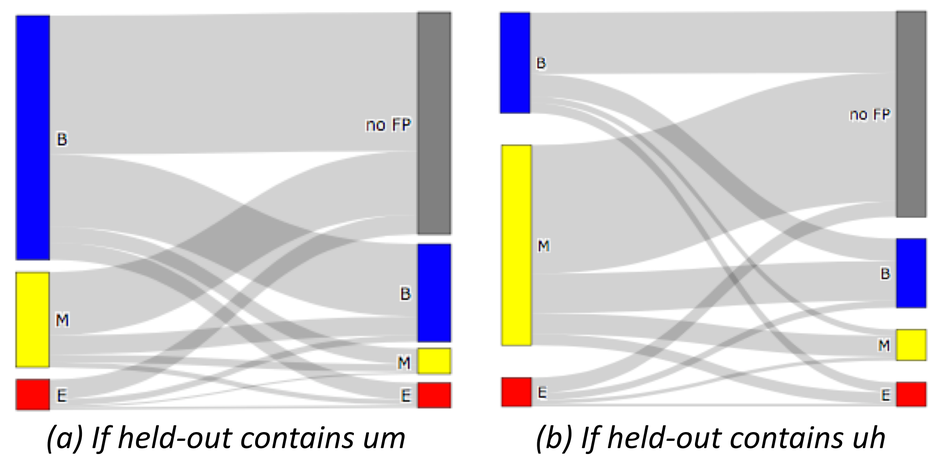
Summary of the voice configurations and the conditions used in the evaluations

System	Corpus & training	Annotation of FPs	Condition	Prompt	Resulting speech
AutoFP	whole TCC	no	**AutoFP**	fluent	has automatically placed FPs
CtrlFP	whole TCC	yes, differentiating ‘uh’ and ‘um’	**CtrlFP-GT**	FPs copied from GT	FPs exactly as in the prompt
			CtrlFP-SW	FPs opposite type as GT	FPs exactly as in the prompt
			CtrlFP-FL	fluent	no FPs
GenFP	whole TCC	yes, with a generic FP label for both ‘uh’ and ‘um’	**GenFP**	Ground-truth FP locations, unspecified type	has FPs in specified locations, type is decided automatically
HalfFluent	fluent 44.4% of TCC	N/A (no FPs in the training data)	**HalfFluent**	fluent	no FPs
TransFluent	whole TCC, then transfer learning to fluent 44.4%	no	**TransFluent**	fluent	very occasional automatically placed FPs

Objective evaluation of automatic FP insertion

FP at *B*	*M*	*E*	Held-out	AutoFP	*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

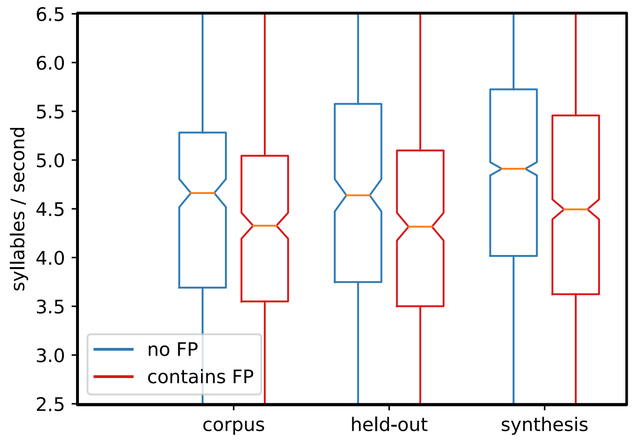
*Breath groups from held-out data with FPs at **B**eginning, **M**iddle, and/or **E**nd*



(a) If held-out contains um

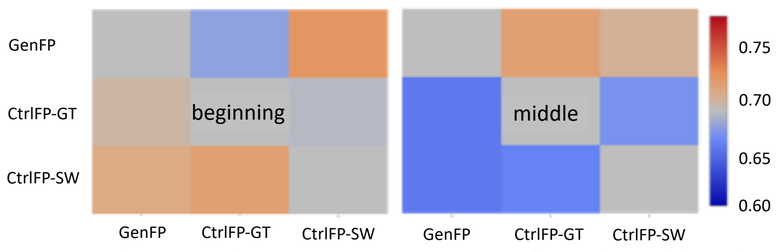
(b) If held-out contains uh

Sankey diagrams of FP position in the held-out data (left in each diagram) and the synthesis (right)



Speech rates within the TCC corpus, the held-out and the synthesised samples, split on whether or not they contain FPs

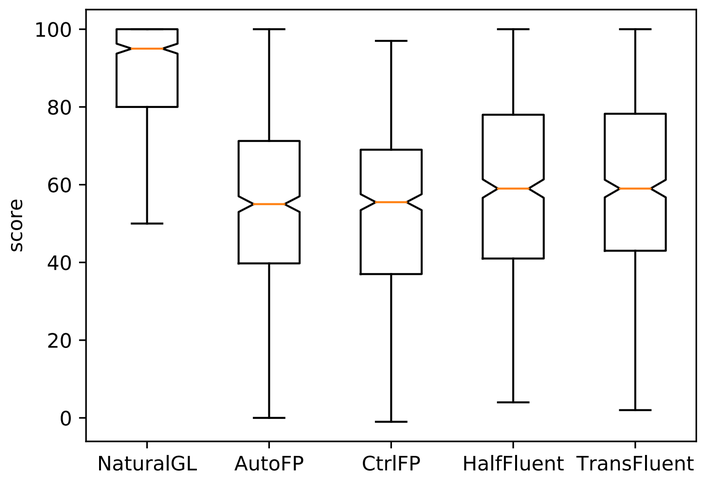
Perceptual evaluation of disfluent speech



Ratio where FP from voice y was rated as plausible when compared with voice x, incl. both being rated plausible

20 utterances from AMI corpus, each containing one ’uh’ or ’um’, in the beginning or the middle.
Pairwise comparison: listeners indicated which version hesitated more realistically.

Perceptual evaluation of fluent speech



MUSHRA ratings of fluent speech

Bonus question

Would you want a robot to sound hesitant? Why?

Yes – 45% “Yes, sounds more authentic and genuine.”
“Yes so it sounds more like a person and more relatable.”
“Yes, much more easy to listen to for prolonged periods.”
“I think it's comforting to have a hesitant voice from them.”

Undecided – 18% “I have no idea.” “Indifferent.”

No – 36% No, because it would sound too human like. Over the phone, I wouldn't be able to tell I am talking to a robot.”
“No, I feel uncomfortable blurring the lines between what sounds naturally human and what is machine.”
“No as I would want it to speak correctly at all times."

Conclusions:
- ✓ *Systems trained with no, or location-only FP annotation reproduce FPs in a similar pattern as in the corpus.*
- ✓ *Synthesiser-predicted FP types (‘uh’ or ‘um’) were preferred over specifying the ground-truth type.*
- ✓ *Using precise annotations and focusing on more fluent parts of the corpus improves naturalness of fluent TTS.*

Listen here! ->

www.speech.kth.se/tts-demos

[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.

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.