OFF THE CUFF: EXPLORING EXTEMPORANEOUS SPEECH DELIVERY WITH TTS

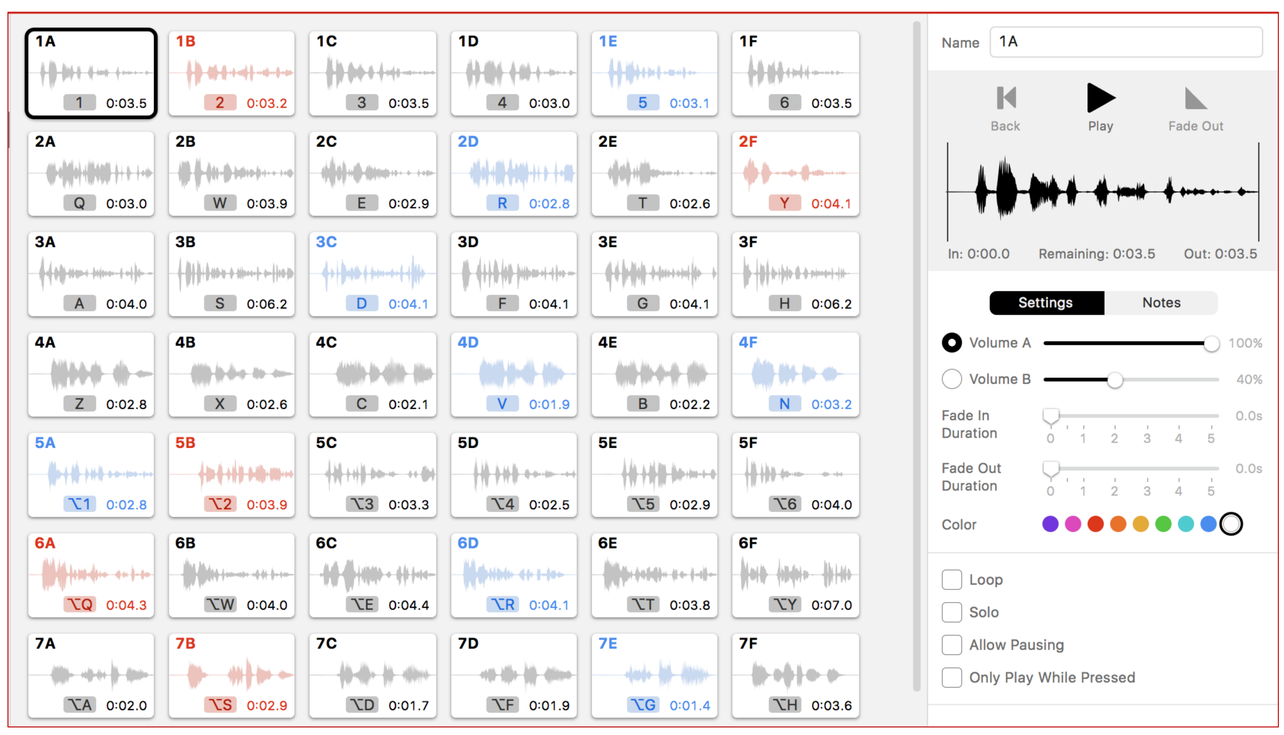
Éva Székely, Gustav Eje Henter, Jonas Beskow, Joakim Gustafson

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

In this demo we present an exploratory platform for interacting with synthetic speech samples speaking part of a public speech. The samples are produced by a spontaneous speech synthesiser and differ in aspects such as fluency and filled pause (FP) placement. Users can colour each sample to mark their subjective impression of how it sounds in the particular in context. This allows investigating context-dependent nuances in speech style such as certainty, authenticity, confidence, etc.

Delivery types in public speaking:

- **Impromptu speaking**: spontaneous speech with no preparation at all
- **Manuscript style:** reading a speech word-for-word
- **Memorised delivery**: committing the entire speech to memory
- **Extemporaneous speech delivery***:* using a structured outline but otherwise delivering the speech conversationally, off the cuff. In this style, the material is presented freely, allowing the speaker to spontaneously change their speech based on listeners' feedback.



Screenshot of the synthetic speech samples on an interactive grid interface of the soundboard app Farrago.

1. Prepare your speech.
2. Synthesise different versions of each utterance.
3. Colour the samples according to your goals, and your subjective impression of what they sound like in context.
4. Deliver your speech to the audience, choosing between styles on-the-go.

#	Voice	Corpus & training	Transcription of FPs	Prompt	Resulting speech
1	AutoFP	whole TCC	no	fluent	has automatically placed FPs
2	CtrlFP	whole TCC	yes, differentiating 'uh' and 'um'	FPs copied from ground truth	FPs exactly as in the prompt
3				fluent	no FPs
4	GenFP	whole TCC	yes, with a generic FP label for both 'uh' and 'um'	Ground-truth FP locations, unspecified type	has FPs in specified locations, type is decided automatically
5	HalfFluent	fluent 44.4% of TCC	N/A (no FPs in the training data)	fluent	no FPs
6	TransFluent	whole TCC, then transfer learning to fluent 44.4%	no	fluent	very occasional automatically placed FPs

Synthetic voice:

ThinkComputers Corpus (TCC): weekly podcast, spontaneous conversational speech, 9h, segmented to single-speaker breath groups, AE, male

TTS: Tacotron + Griffin-Lim

Potential application areas:

- *Communication aids*
- *Rehearsing public speeches*
- *Language learning*
- *TTS evaluation*

For further details on the TTS , please refer to:

[1] É. Székely, G. E. Henter, J. Beskow and J. Gustafson, "Spontaneous conversational speech synthesis from found data" in *Proc. Interspeech*, 2019.

[2] É. Székely, G. E. Henter, J. Beskow and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis" in *Proc. 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.