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

Abstract

Extemporaneous speech is a delivery type in public speaking which uses a structured outline but is otherwise delivered conversationally, off the cuff. This demo uses a natural-sounding spontaneous conversational speech synthesiser to simulate this delivery style. We resynthesised the beginnings of two Inter-speech keynote speeches with TTS that produces multiple different versions of each utterance that vary in fluency and filled-pause placement. The platform allows the user to mark the samples according to any perceptual aspect of interest, such as certainty, authenticity, confidence, etc. During the speech delivery, they can decide on the fly which realisation to play, addressing their audience in a connected, conversational fashion. Our aim is to use this platform to explore speech synthesis evaluation options from a production perspective and in situational contexts.

Index Terms: Spontaneous speech synthesis, public speaking, speech synthesis evaluation, filled pauses, AAC, soundboard

1. Introduction

Public speaking styles can be ordered in four categories according to delivery [1]: impromptu speaking involves spontaneous speech with no preparation at all; manuscript style is reading a speech word-for-word from its written form, while memorised delivery means committing the entire speech to memory. Extemporaneous public speaking, on the other hand, is done on the basis of a prepared structure, such as notes or an outline, but is otherwise delivered off the cuff. In this style, the material is presented freely, allowing the speaker to change their speech based on listeners’ feedback. Communication coaches often advice against manuscript or memorised delivery styles unless the situation is very formal, because of the risk of sounding “robotic”. It is no surprise that robots are used as a reference as conventionally, speech synthesisers (even expressive ones) deliver spoken material by reading text out loud, based on speech data from people doing the same. However, with the rise of natural-sounding spontaneous speech synthesis, TTS versions is rather context-dependent.

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 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. Our hope is that this application will be a discussion starter in the scientific community about the need to move away from isolated utterances in TTS evaluation and the options of evaluating synthesis from the production perspective, where subjects interact with and use the TTS to attain their communicative goals.

2. Spontaneous speech synthesis

2.1. Spontaneous speech data and TTS

The data we use in this study is an untranscribed weekly technology podcast, called the “ThinkComputers” podcast, available in the public domain via the Internet Archive (archive.org). The recordings contain product reviews and discussions of technology news from two male speakers of American English mixed into a single audio channel. To segment the data into clean, well-defined utterances we used the speaker-dependent breath detection method proposed in [2]. With this method we selected 6,218 speech segments from 27 podcast episodes, each starting with a breath event from the target speaker. The utterances, henceforth referred to as the ThinkComputers Corpus (TCC), were automatically transcribed using the Google Cloud Speech API [3]. Filled pauses (FPs) were identified using the Gentle forced aligner [4]. For further details please refer to [5].

All versions of the voice described here were built using the implementation [6] of the Tacotron 2 spectrogram prediction framework [7]. All audio was sampled at 22.1 kHz. The Griffin-Lim algorithm [8] was used for waveform synthesis. Samples from this voice can be found under www.speech.kth.se/tts-demos.

2.2. Voice variants

We have produced several different speech variants – summarised in Table 1 – with differences in how disfluencies were addressed during annotation, training, and synthesis. We have previously [5] found our TCC voices to be rated significantly more appropriate than read speech synthesis on prompts from public speeches. However, the perceptual effect of the different speech variants is rather context-dependent.

Speech version 1 is synthesised by a voice we call AutoFP. This was built using the entire TCC corpus, but filled pauses ‘uh’ and ‘um’ were not annotated. This resulted in the synthes-
Table 1: Summary of the six different synthetic speech versions used in our demo.

<table>
<thead>
<tr>
<th>Version</th>
<th>Name of voice</th>
<th>Corpus and training</th>
<th>Transcription of FPs</th>
<th>Prompt</th>
<th>Resulting speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AutoFP</td>
<td>whole TCC</td>
<td>no</td>
<td>fluent</td>
<td>has automatically placed FPs</td>
</tr>
<tr>
<td>2</td>
<td>CtrlFP</td>
<td>whole TCC</td>
<td>yes, differentiating ‘uh’ and ‘um’</td>
<td>fluent</td>
<td>FPs copied from ground truth FPs exactly as in the prompt</td>
</tr>
<tr>
<td>3</td>
<td>GenFP</td>
<td>whole TCC</td>
<td>yes, with a generic FP label for both ‘uh’ and ‘um’</td>
<td>ground-truth FP locations, unspecified type</td>
<td>has FPs in specified locations, type is decided automatically</td>
</tr>
<tr>
<td>4</td>
<td>HalfFluent</td>
<td>fluent 44.4% of TCC</td>
<td>N/A (no FPs in the training data)</td>
<td>fluent</td>
<td>60 FPs</td>
</tr>
<tr>
<td>5</td>
<td>TransFluent</td>
<td>whole TCC, then transfer learning to the fluent 44.4%</td>
<td>no</td>
<td>fluent</td>
<td>very occasional automatically placed FPs</td>
</tr>
</tbody>
</table>

A sample task, presented in the video attached to this demo paper, involves colour-tagging randomly ordered versions of each text prompt, such that a clicking through a progression of same-coloured utterances gives an impression of either a confident or uncertain speaker (coloured blue and red in Figure 1).

4. Use case scenarios for simulating extemporaneous speech production

Our proposed platform can be used to develop TTS evaluation strategies that take a production perspective, and also allows for perceptual judgments to take place in specific situational contexts. Another use case scenario is an application for Alternative and Augmentative Communication (AAC), specifically for people who use synthetic speech as their main verbal communication method because of a medical condition. The user can prepare a speech by synthesising different versions of each sentence in their presentation, then rehearse the speech in a similar manner as people who speak with their natural voice. The platform would allow samples to be annotated by the users according to any communicative nuances they find relevant. This would let AAC users modify their speech delivery on the fly, enabling them to be responsive to their audience in the moment.

5. Acknowledgements

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).

6. References