Context-Aware Prosody Correction for Text-Based Speech Editing

Max Morrison, Lucas Rencker, Zeyu Jin, Nicholas J. Bryan, Juan-Pablo Caceres, Bryan Pardo

IEEE ICASSP 2021

Overview

We propose a method for correcting the mismatch in prosody caused by directly copying and pasting speech waveforms. Consider replacing the words highlighted in blue in Figure 1 with the phrase highlighted in purple in Figure 2 by directly copying and pasting the waveform (and corresponding pitch). The resulting naïvely copy-pasted speech (Figure 3) sounds unnatural, as the prosody of the pasted speech does not match the context. We correct the naïvely copy-pasted speech with our proposed method (Figure 5). The result is shown in Figure 4.

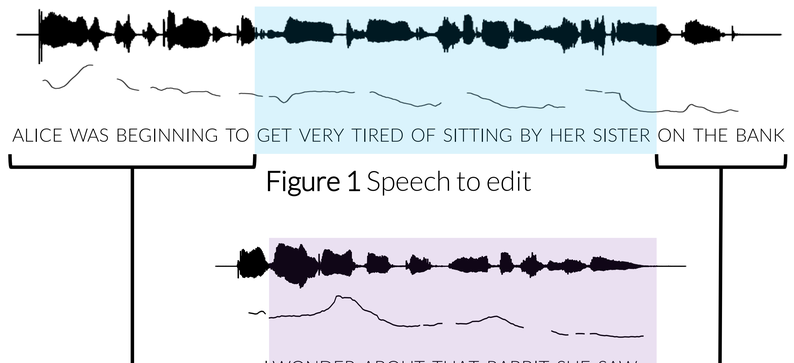

Figure 1 Speech to edit

Figure 2 Speech with desired phrase

Figure 3 Naïve copy-paste (unnatural prosody)

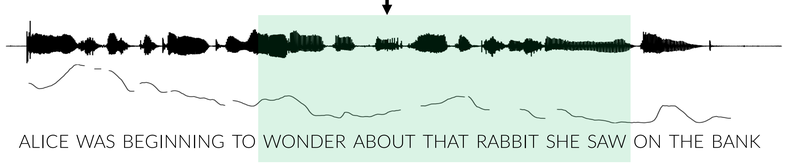

Figure 4 After prosody correction

Listen to audio examples at
maxrmorrison.com/sites/context-aware

Method

Inputs – naïvely copy-pasted speech (e.g., Figure 3) and corresponding phonemes

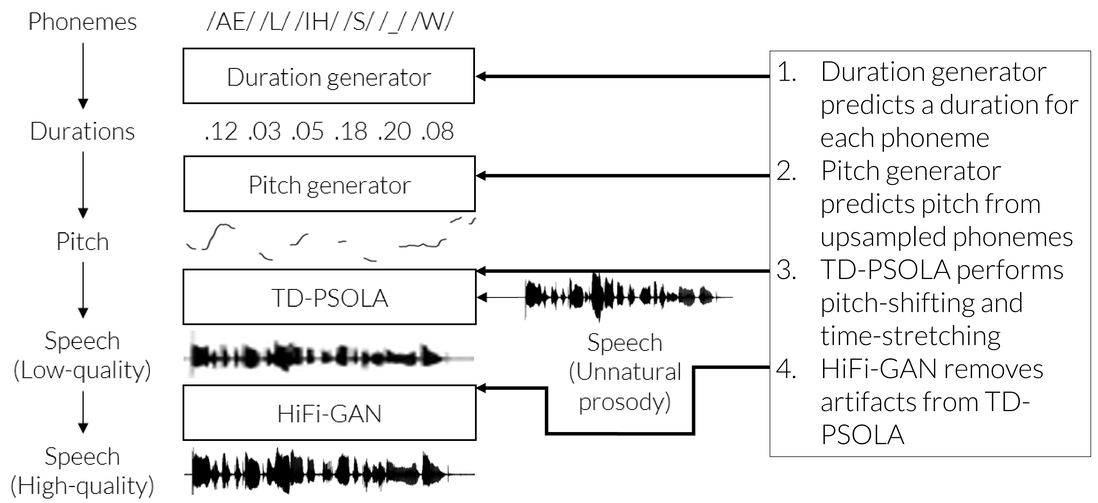

Figure 5 Our proposed method for prosody correction

Evaluation

We conduct naturalness tests using mean opinion score (**MOS**), with range 1 (worst) to 5 (best), and a **Pairwise** test, which measures percent preference for our system. We compare **Original** speech, **Naïvely** copy-pasted speech, speech with **Average** pitch and durations, our **Proposed** method, four ablations, and speech with prosody from a **Tacotron**-based model.

Method	MOS	Pairwise
Original	4.56	8.56%
Naive	2.87	57.7%
Average	2.76	62.7%
Proposed	**3.00**	-
-Duration	2.97	45.3%
-Pitch	2.86	55.8%
-HiFi-GAN	2.87	54.2%
-Context	2.59	67.9%
Tacotron	2.89	53.2%

Conclusions

1. Context-awareness is important for prosody generation
2. Our method performs more natural copy-and-paste of speech
3. Neural speech enhancement can improve quality of DSP-based speech manipulation