Context-Aware Prosody Correction for Text-Based Speech Editing

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

*Northwestern University  †University of Surrey  †Adobe Research

IEEE ICASSP 2021
Context-Aware Prosody Correction

• Naively copy-pasting speech waveforms sounds unnatural
• Make it sound natural by changing the prosody to match the context
• Permits text-based speech editing with natural prosody
Speech editing via prosody correction

Speech to edit

Sentence with desired phrase

Naïve copy-paste

After prosody correction
Speech editing via prosody correction

ALICE WAS BEGINNING TO GET VERY TIRED OF SITTING BY HER SISTER ON THE BANK
ALICE WAS BEGINNING TO WONDER ABOUT THAT RABBIT SHE SAW ON THE BANK
Speech editing via prosody correction

ALICE WAS BEGINNING TO WONDER ABOUT THAT RABBIT SHE SAW ON THE BANK

Neural network
Outline

- Methods
- Evaluation
- Conclusion
Prosody correction (proposed)

Phonemes

Durations

Pitch

Speech (Low-quality)

Speech (High-quality)

//AE// /L/ /IH/ /S/ // /N//

Duration generator

.12 .03 .05 .18 .20 .08

Pitch generator

TD-PSOLA

HiFi-GAN

Speech (Unnatural prosody)

Related prior work
• FastSpeech 2 [Ren et. al., 2020]
  • Generates pitch and phoneme duration from text
  • Not context-aware
  • Not multi-speaker
Phoneme duration generation (proposed)

Listen for the words “the number of visitors”

Ground truth

0.12 0.03 0.05 0.18 0.20 0.08

MSE Loss

Durations

0.12 ? ? ? ? 0.08

Unidirectional GRU

Bidirectional GRU

1D Convolution + ReLU

x2

Phonemes

/AE/ /L/ /IH/ /S/ /I/ /W/
Pitch generation with C-DAR

Listen for the words “on the floor”

Ground truth

Cross Entropy Loss

Pitch

1D Convolution + ReLU  \times 5

Autoregressive GRU

Bidirectional GRU

1D Convolution + ReLU  \times 5

Upsample

Phonemes and durations

/AE/ /L/ /IH/ /S/ /_/ /W/
\quad 0.12 \ 0.03 \ 0.05 \ 0.18 \ 0.20 \ 0.08

Related prior work
- Autoregressive [Wang et. al., 2018]
- VQVAE-based [Wang et. al., 2019]

[Morrison et. al., 2020]
Pitch-shifting and time-stretching with TD-PSOLA

Related prior work
- STRAIGHT [Banno et. al., 2007]*
- WORLD [Morise et. al., 2016]*
- TD-PSOLA [Moulines and Charpentier, 1990]

* WORLD has been shown to have much higher subjective quality than STRAIGHT [Morise and Watanabe, 2018]

GPL-licensed code for TD-PSOLA available at [github.com/maxrmorrison/psola]

[Moulines and Charpentier, 1990]
Denoising with HiFi-GAN

- HiFi-GAN [Su et. al., 2020] is a neural speech enhancement method
- Adversarial loss promotes generalization to various types of noise, reverb, and artifacts
- Can HiFi-GAN improve the quality of TD-PSOLA?
Outline

- Methods
- Evaluation
- Conclusion
Speech replacement task

Listen for the words “was about”

Speech to edit

Naïve copy-paste

Same words in different sentence

After prosody correction
Results

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Naive</th>
<th>Average</th>
<th>Proposed</th>
<th>-Duration</th>
<th>-Pitch</th>
<th>-HiFi-GAN</th>
<th>-Context</th>
<th>Tacotron</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOS</td>
<td>4.56</td>
<td>2.87</td>
<td>2.76</td>
<td>3.00</td>
<td>2.97</td>
<td>2.86</td>
<td>2.87</td>
<td>2.59</td>
<td>2.89</td>
</tr>
<tr>
<td>Pairwise</td>
<td>8.56%</td>
<td>57.7%</td>
<td>62.7%</td>
<td>-</td>
<td>45.3%</td>
<td>55.8%</td>
<td>54.2%</td>
<td>67.9%</td>
<td>53.2%</td>
</tr>
</tbody>
</table>

- Most significant components
  - Context-awareness
  - Pitch regeneration
  - HiFi-GAN

- Future work
  - Speech manipulation quality
  - Multi-speaker duration modeling
Conclusions

1. Context-awareness is important for prosody generation
2. We propose a method for context-aware prosody generation
3. Listening tests show our method performs more natural copy-and-paste of speech
4. Neural speech enhancement can improve quality of DSP-based speech manipulation

Hear more audio examples at maxrmorrison.com/sites/context-aware.

Thanks for listening!