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

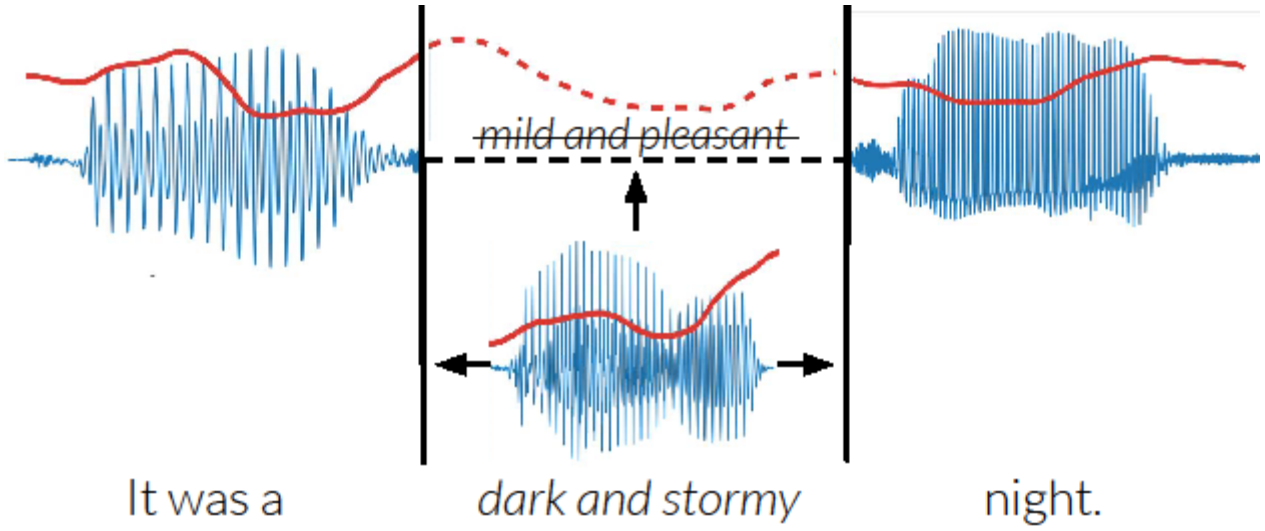
‡University of Surrey

†Adobe Research

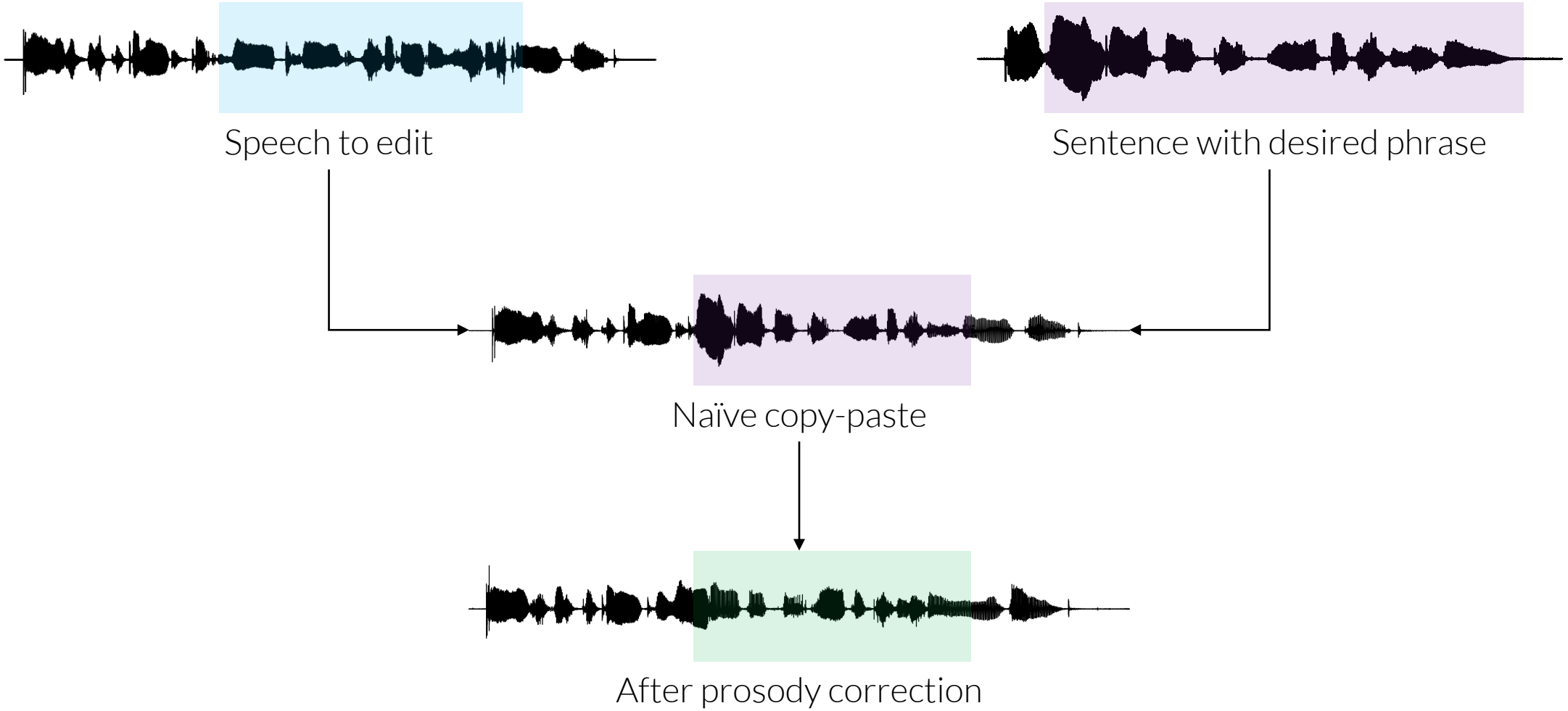
IEEE ICASSP 2021

Context-Aware Prosody Correction

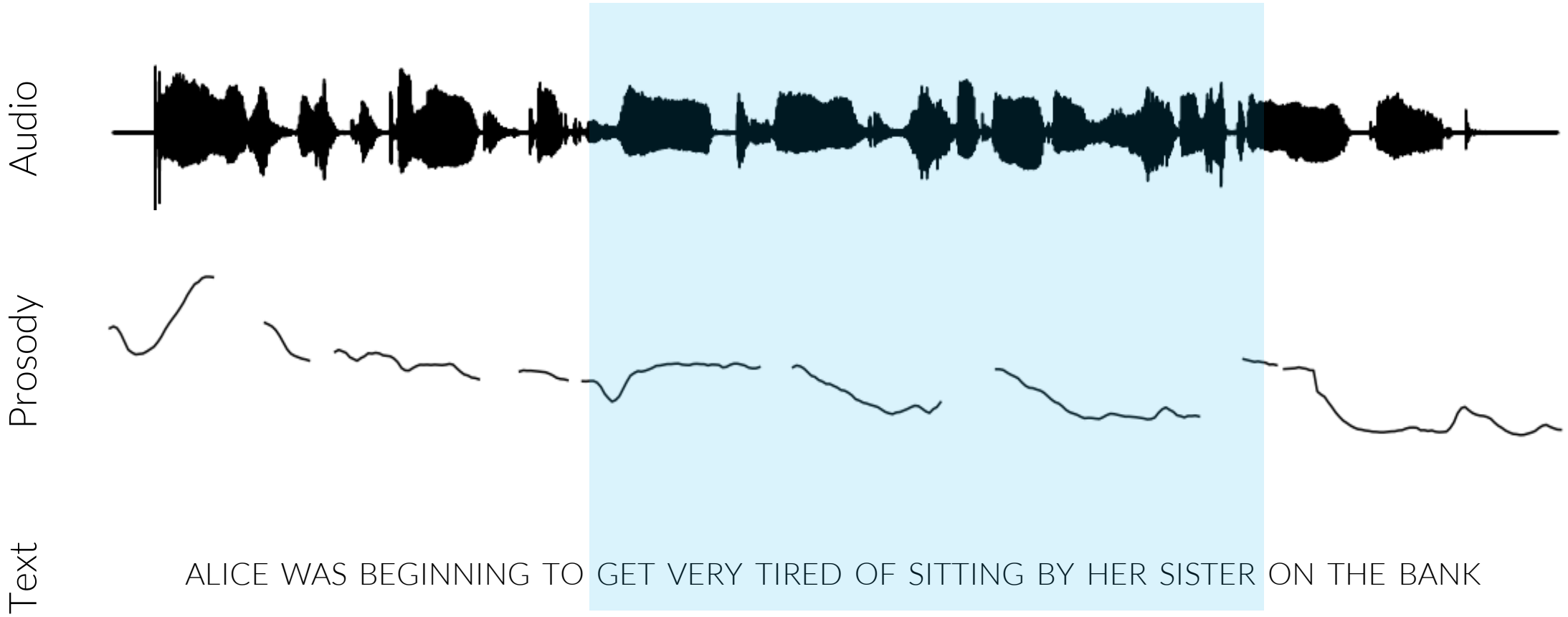
- Naively copy-pasting speech waveforms sounds unnatural
- Make it sound natural by changing the prosody to match the context
- Permits text-based speech editing with natural prosody



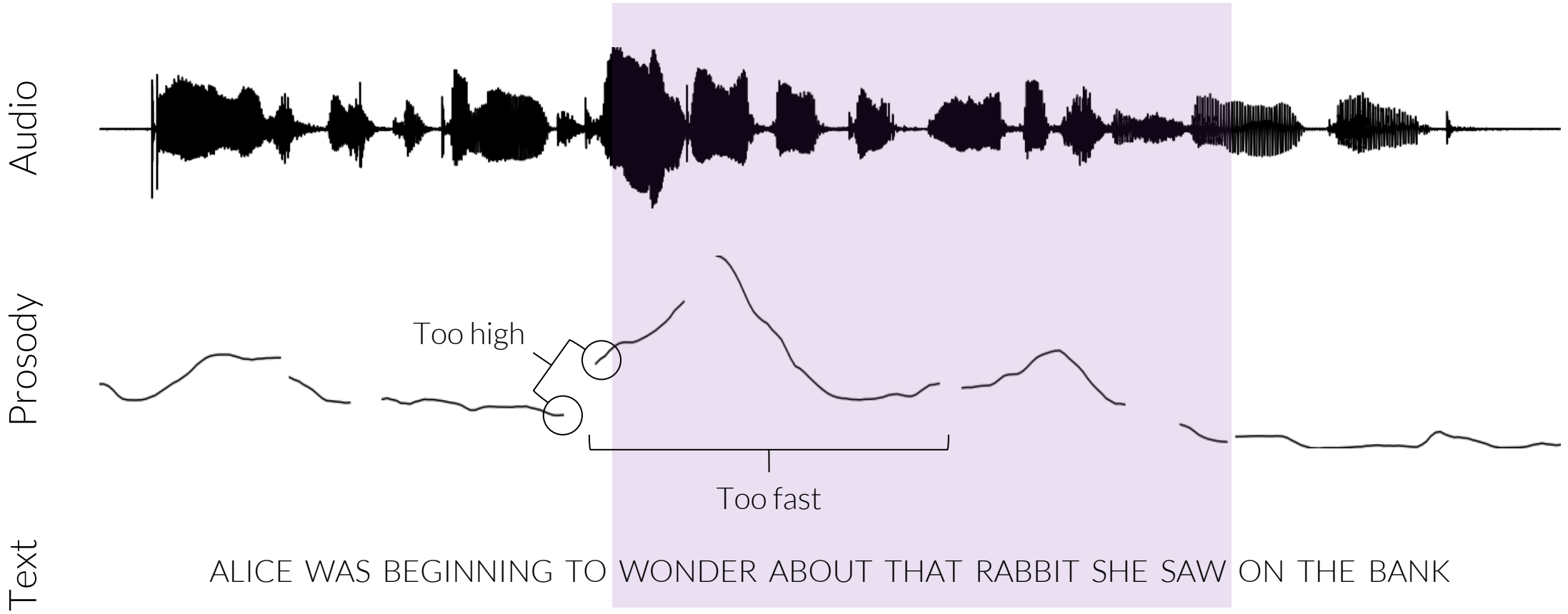
Speech editing via prosody correction



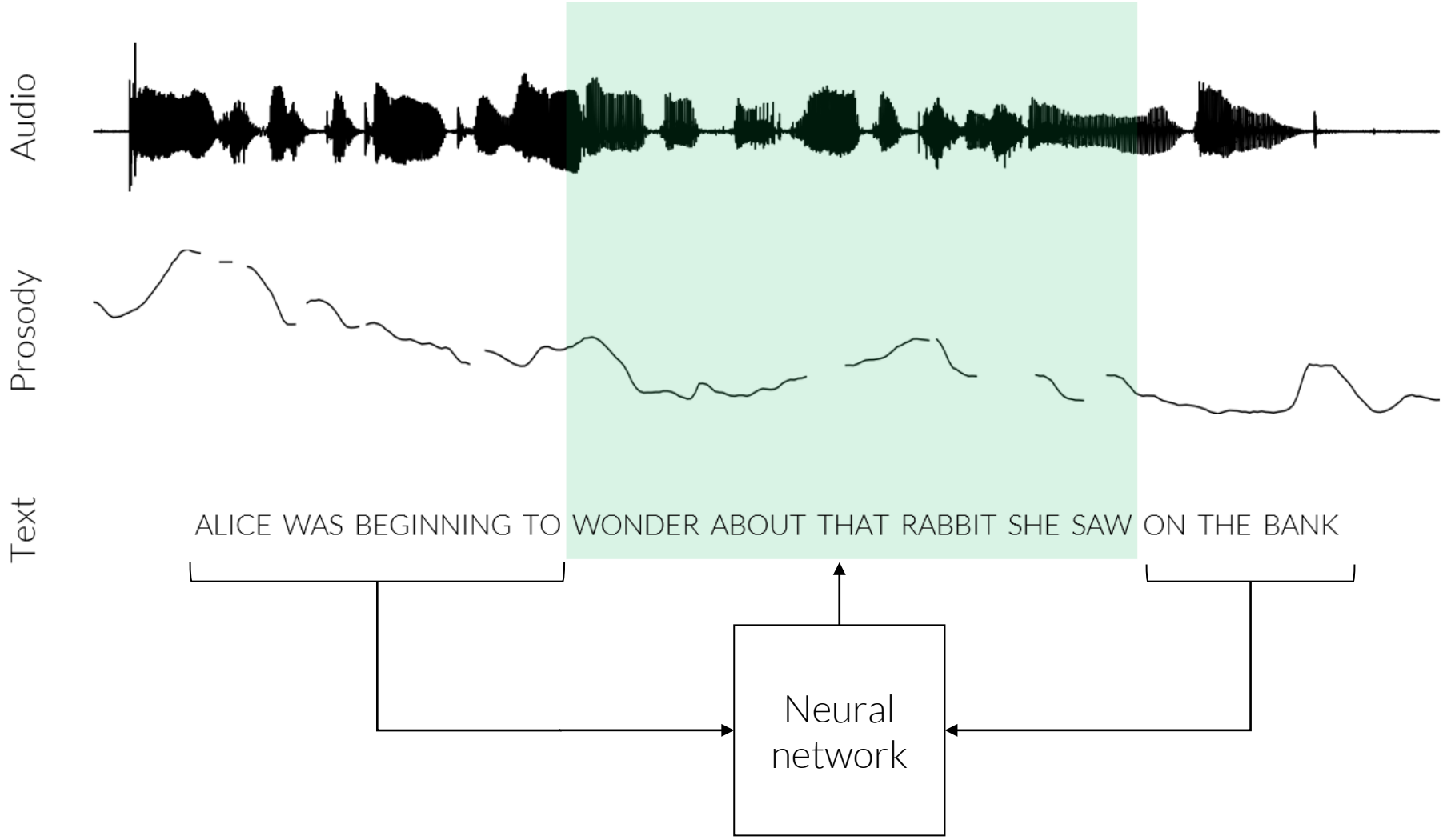
Speech editing via prosody correction



Speech editing via prosody correction



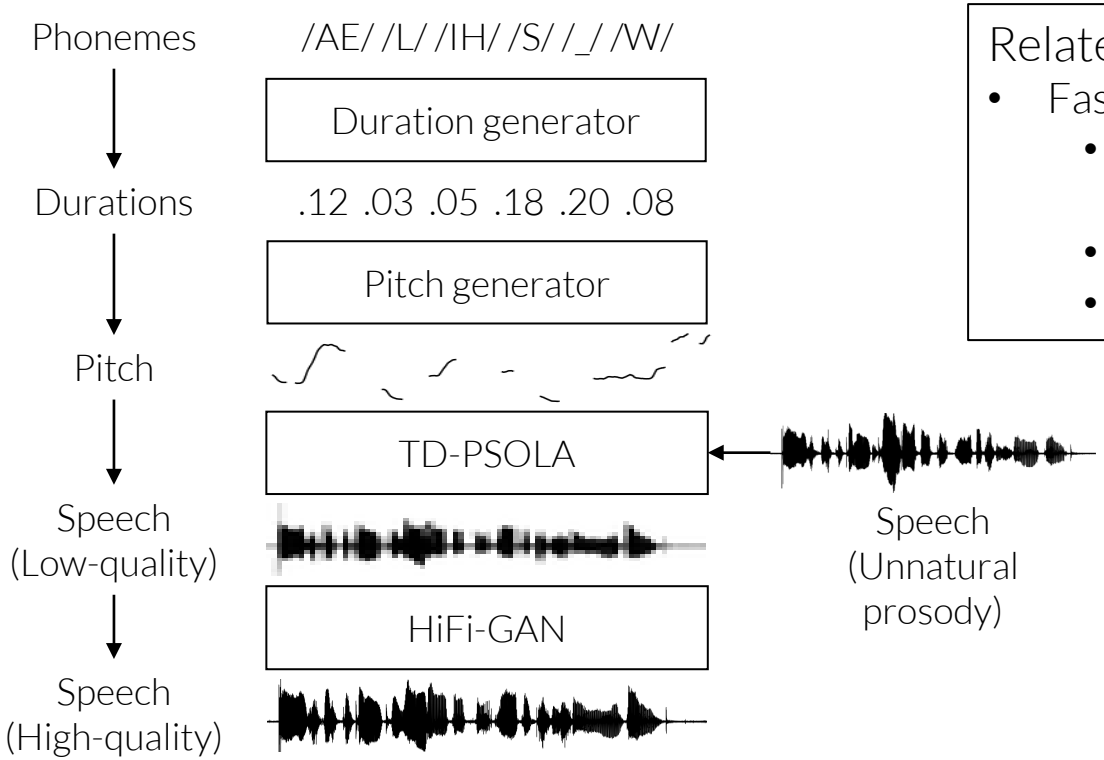
Speech editing via prosody correction



Outline

- Methods
- Evaluation
- Conclusion

Prosody correction (proposed)

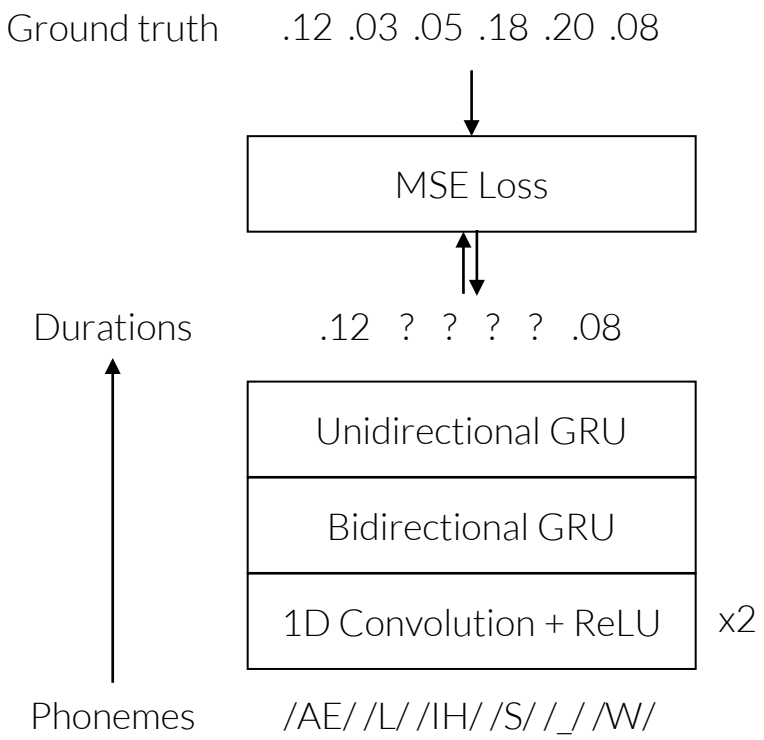


Related prior work

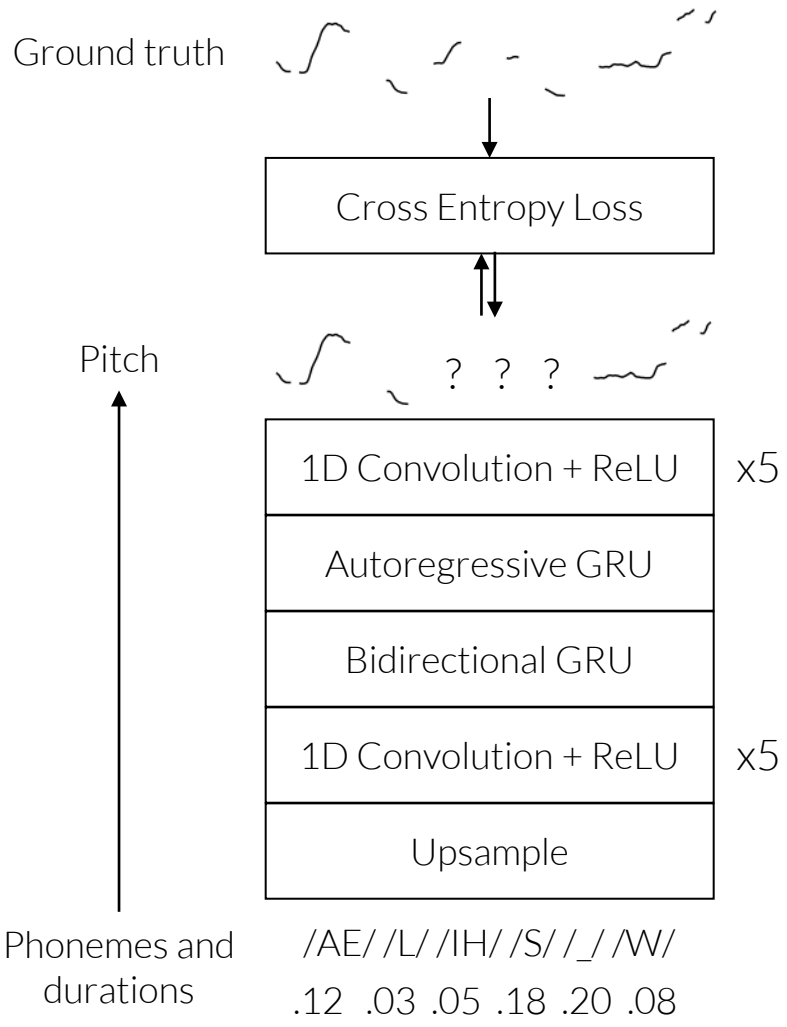
- FastSpeech 2 [Ren et. al., 2020]
  - Generates pitch and phoneme duration from text
  - Not context-aware
  - Not multi-speaker

Phoneme duration generation (proposed)

Listen for the words
“the number of visitors”



Pitch generation with C-DAR



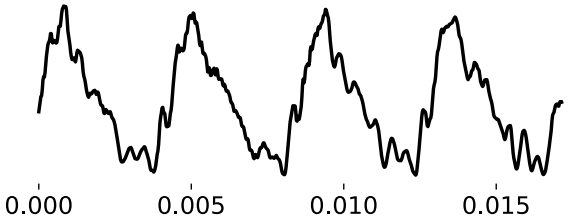
Related prior work
- Autoregressive [Wang et. al., 2018]
- VQVAE-based [Wang et. al., 2019]

Listen for the words "on the floor"

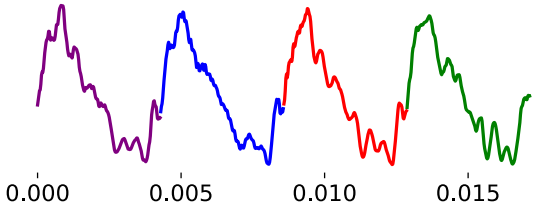
[Morrison et. al., 2020]

Pitch-shifting and time-stretching with TD-PSOLA

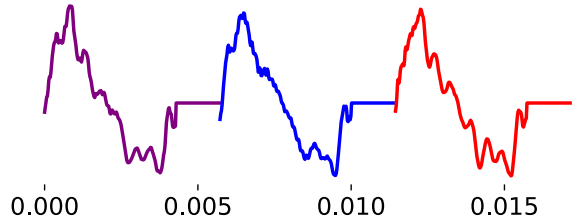
Related prior work
- STRAIGHT [Banno et. al., 2007]*
- WORLD [Morise et. al., 2016]*
- TD-PSOLA [Moulines and Charpentier, 1990]



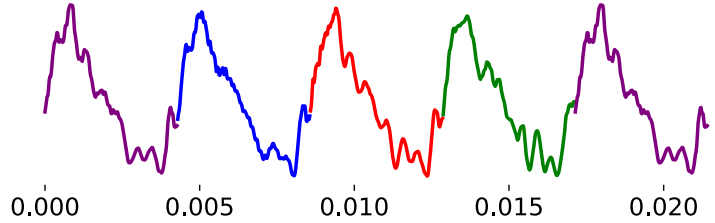
Original speech



Identify repeating waveforms



Pitch-shift via decomposition



Time-stretch via repetition

* 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]

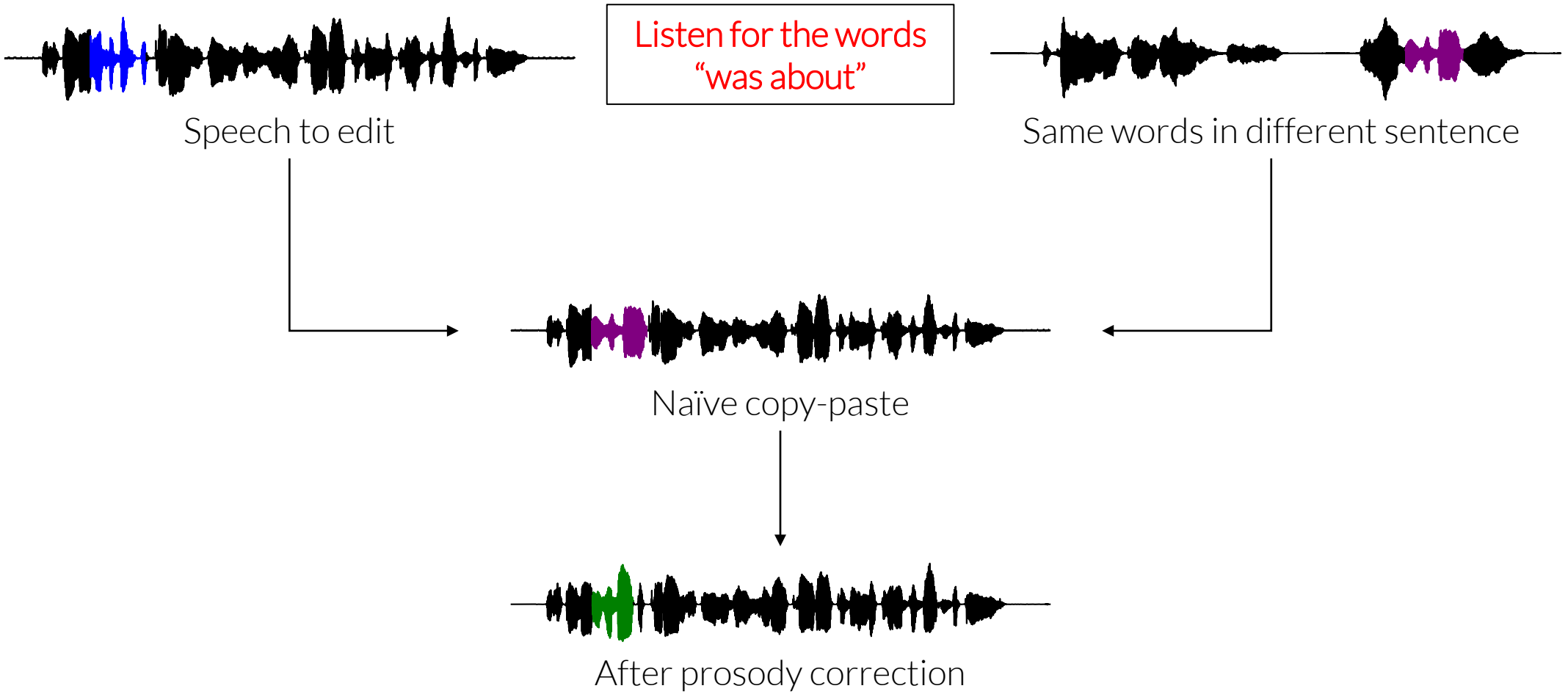
Denoising with HiFi-GAN

- 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



Results

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

- Most significant components
  - Context-awareness
  - Pitch regeneration
  - HiFi-GAN
- Future work
  - Speech manipulation quality
  - Multi-speaker duration modeling

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

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!