Speech Denoising by Parametric Resynthesis

Soumi Maiti¹, Michael Mandel²
¹City University of New York, NY-USA, ²Brooklyn College, City University of New York, NY-USA

Motivation
- High-quality speech enhancement
- Utilize WORLD vocoder to synthesize output “clean” speech
- Inspired by parametric speech synthesis

Parametric Resynthesis
Resynthesize clean speech by predicting acoustic parameters

Prediction Model
Predict clean acoustic parameters → noisy speech
- Similar to acoustic modelling in TTS
- Predicts features at a fixed frame rate
  → 46 ms frame, 5 ms hop size
- 2 layer LSTM, 512 units
- Loss: MSE
- Output:
  - Spec Env
  - F0
  - V/UV
  - AP

Vocoder
Vocoder: WORLD¹
- Encode: clean speech → acoustic parameters
  - Generate labels for prediction model
- Decode: acoustic parameter → clean speech
  - Synthesize output “clean” speech

Experiments
- CMU arctic speech dataset¹ → slt: female
- Add CHiME3² environmental noise
  → Bus, café, street, pedestrian
- Train/dev/test: 1000/66/66
- SNR: -6 dB to 21 dB

Systems:
- PR-clean: PR with clean speech → upper bound on PR
- PR: Parametric resynthesis
- TTS: statistical text-to-speech
- DNN-IRM: speech enhancement
- OWM: Oracle Wiener mask → access to clean speech

TTS objective Measures

Listening Test
Randomly selected 12 test files
Quality: MUSHRA test

Intelligibility

Two Speaker
slt: female speech, bdl: male speech

Summary
- Outperforms TTS by capturing prosody
- Outperforms DNN-IRM in listening test
- Comparable to oracle system

References:

Acknowledgement:
This work is supported by the NSF under Grant IIS-1618061. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Demo: http://mr-oc-work/icassp19/

SNR:
- 10.41
- 4.81
- 12.17

PR:
- 4.81
- 4.09
- 12.17

TTS:
- 5.05
- 5.40
- 12.34

PR-clean:
- 2.68
- 4.91
- 10.41

PR:
- 4.81
- 4.91
- 10.41

TTS objective Measures

Listening Test

Intelligibility

Two Speaker

Summary

Same model can be used for multiple speakers