Speech Denoising by Parametric Resynthesis

Soumi Maiti[1], Michael Mandel[1,2]

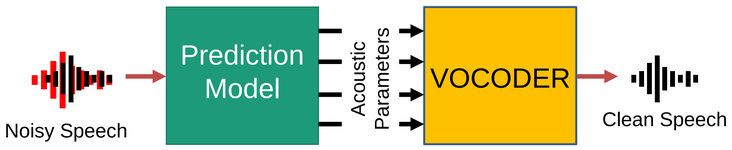
[1]City University of New York, NY-USA, [2]Brooklyn College, City University of New York, NY-USA

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

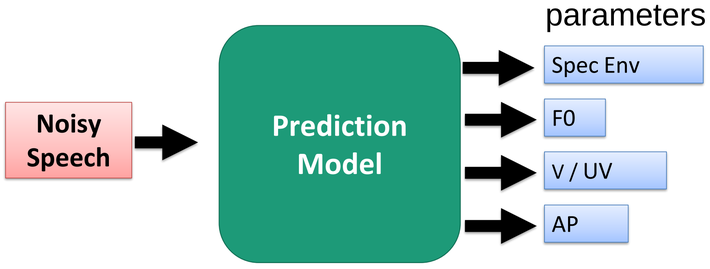


Prediction Model

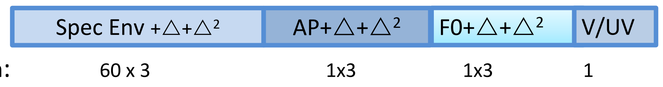
Predict clean acoustic parameters ← noisy speech

- 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 :



Demo: http://mr-pc.org/work/icassp19/

References:

1. M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," IEICE Trans. Info. Sys., 99(7), 1877-1884, 2016.
2. J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chime speech separation and recognition challenge: Dataset, task and baselines," Proc. ASRU, 2015, pp. 504–511.
3. J. Kominek, and A. W. Black, "The CMU Arctic speech databases," Proc. ISCA Wkshp. speech synthesis, 2004.

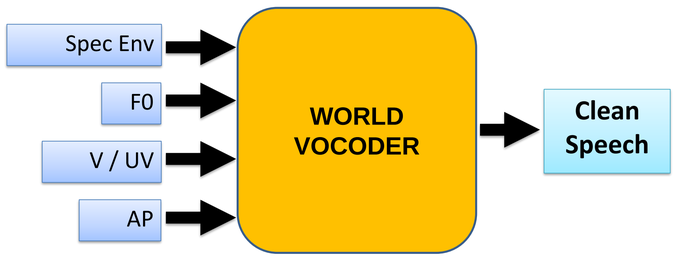
Acknowledgement:

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.

Vocoder

Vocoder: WORLD[1]

- Encode: clean speech → acoustic parameters
  - Generate labels for prediction model
- Decode: acoustic parameter → clean speech
  - Synthesize output "clean" speech



Four acoustic parameters of vocoder:

1. **Spec Env**: spectral envelope
2. **AP**: Aperiodic energy
3. **F0**: log fundamental frequency
4. **V/UV** : voicing decision (0/1)

Experiments

- CMU arctic speech dataset[3] → slt : female
- Add CHiME3[2] 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

	Spectral distortion		F0 measures		
	MCD ↓	BAPD ↓	RMSE ↓	CORR ↑	VUV ↓
PR-clean	2.68	0.16	4.95	0.96	2.78%
TTS	5.05	0.24	12.60	0.73	5.60%
PR	**4.81**	**0.19**	**5.62**	**0.95**	**5.27%**

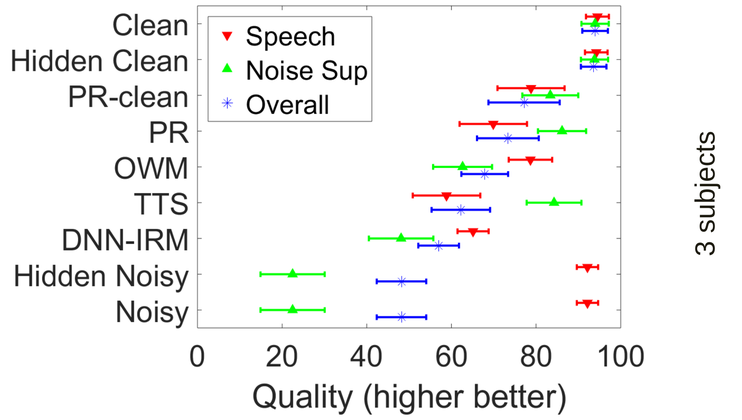
PR outperforms TTS

→ Captures realistic prosody from noisy speech

Listening Test

Randomly selected 12 test files

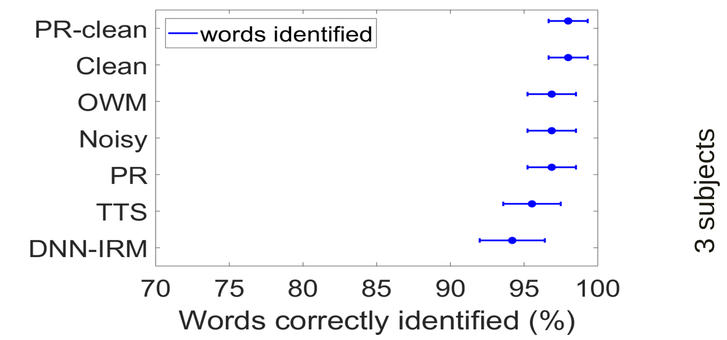
Quality: MUSHRA test



PR outperforms DNN-IRM!

PR equals Oracle Wiener mask!

Intelligibility



High intelligibility!

Two Speaker

slt: female speech, bdl: male speech

		Spectral distortion		F0 measures		
Train	Test	MCD ↓	BAPD ↓	RMSE ↓	CORR ↑	VUV↓
slt	slt	4.81	0.19	5.62	0.95	5.27%
slt+bdl	slt	4.91	0.20	8.36	0.92	6.50%
bdl	bdl	5.40	0.21	9.67	0.82	12.34%
slt+bdl	bdl	5.19	0.21	10.41	0.82	12.17%

Same model can be used for multiple speakers

Summary

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