Speaker independence of neural vocoders and their effect on parametric resynthesis speech enhancement

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

[1]The Graduate Center, City University of New York

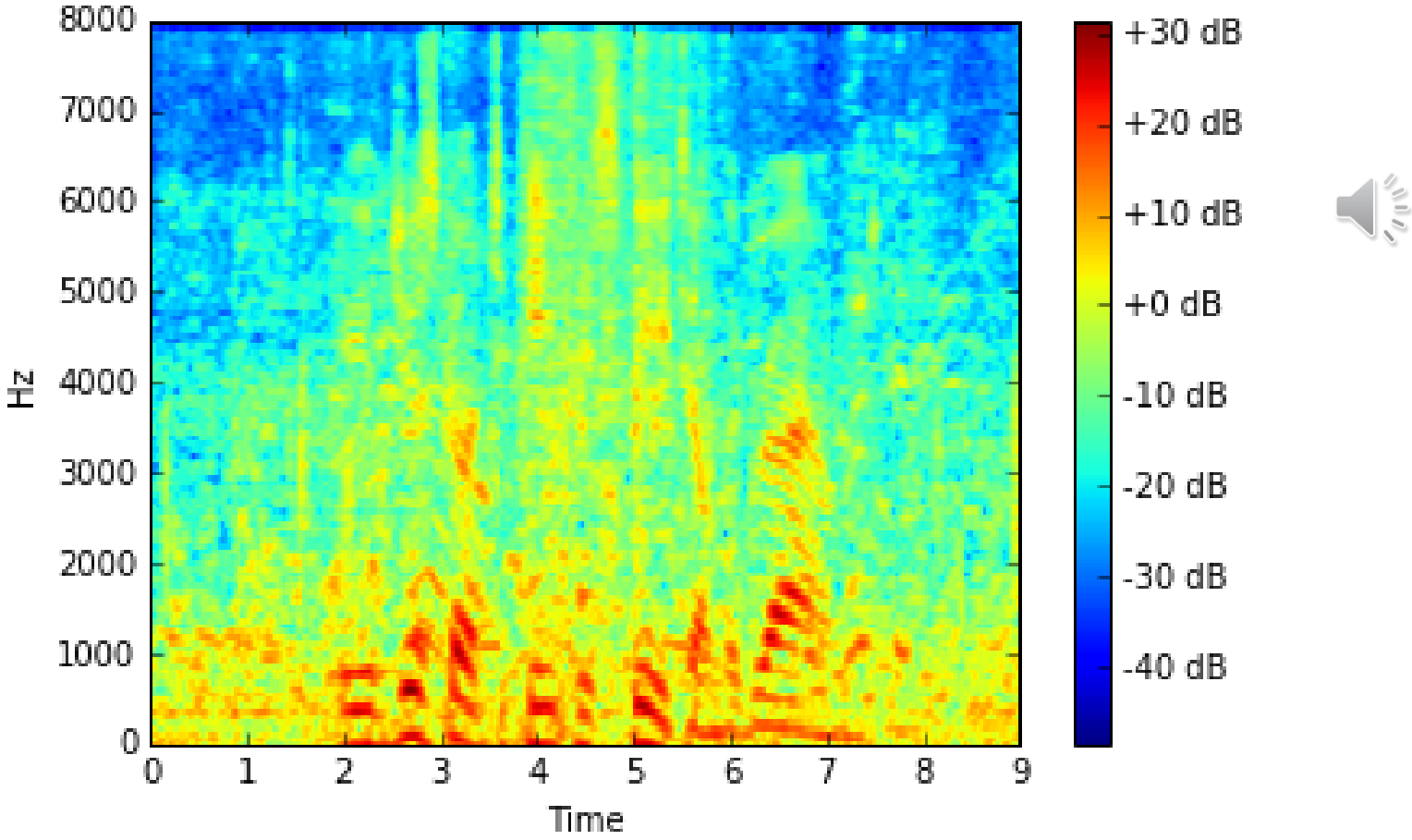
[2]Brooklyn College, City University of New York

ICASSP 2020

THE GRADUATE CENTER
CITY UNIVERSITY OF NEW YORK

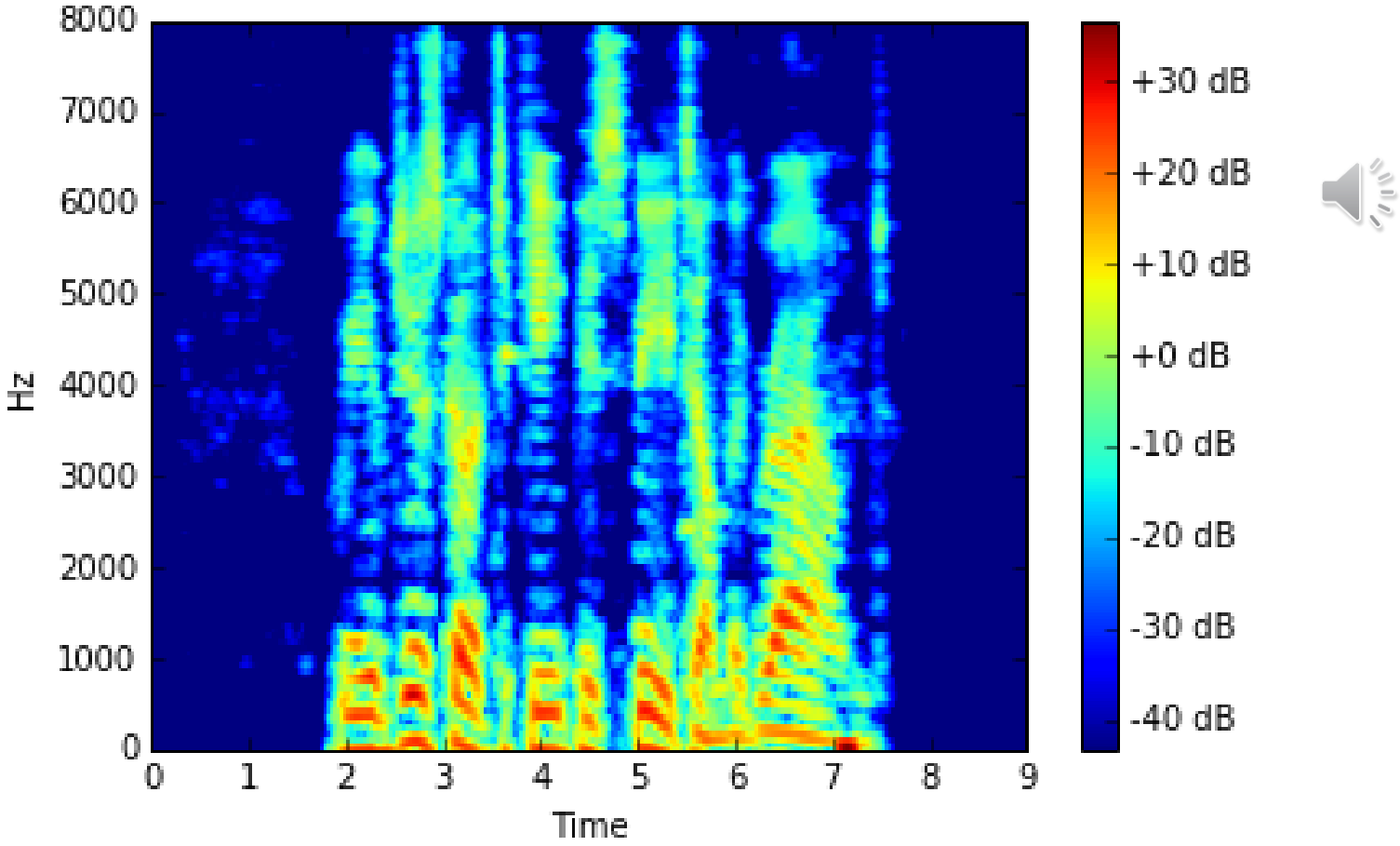
Example: Noisy file

SNR: 2.5 dB
Café noise



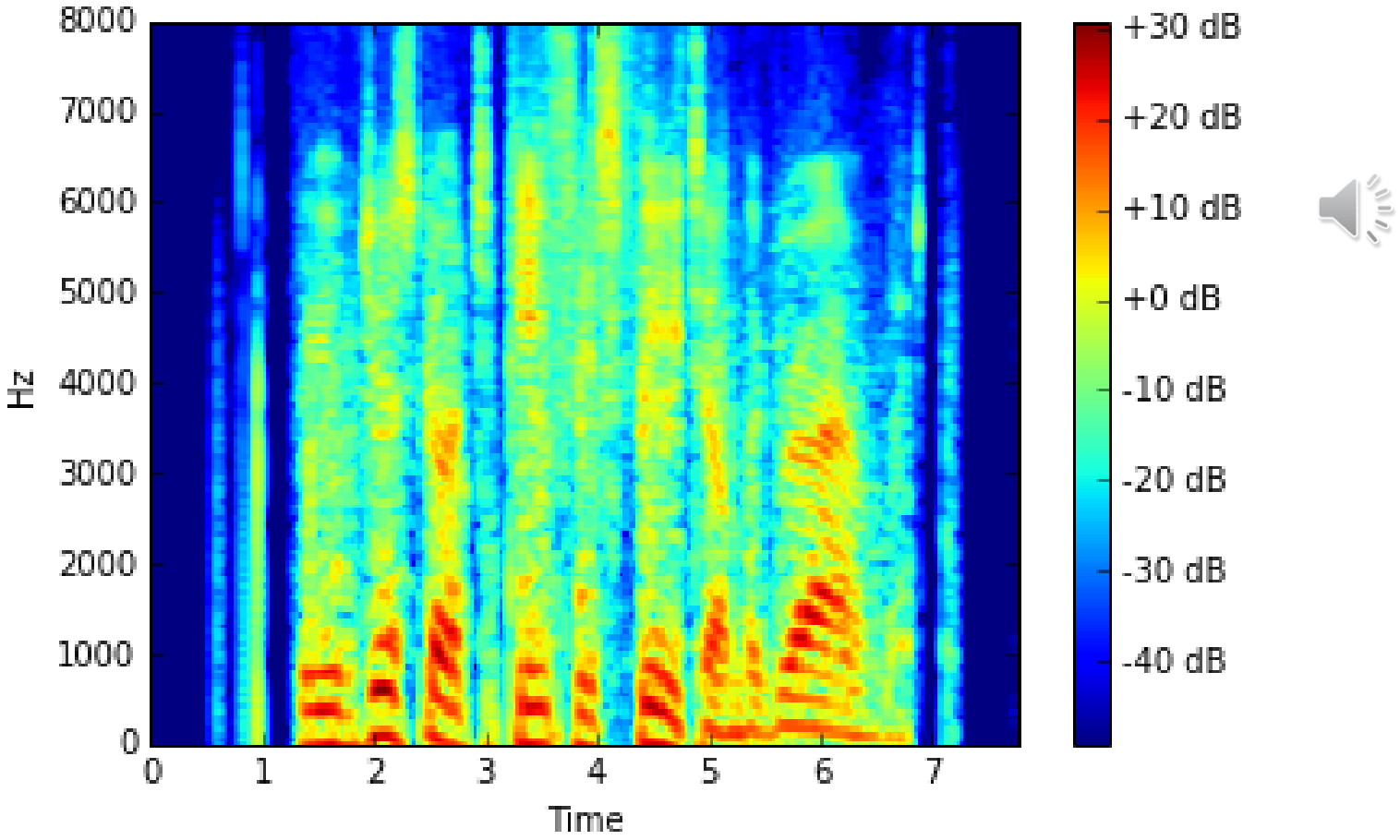
Example: Oracle mask system

SNR: 2.5 dB
Café noise



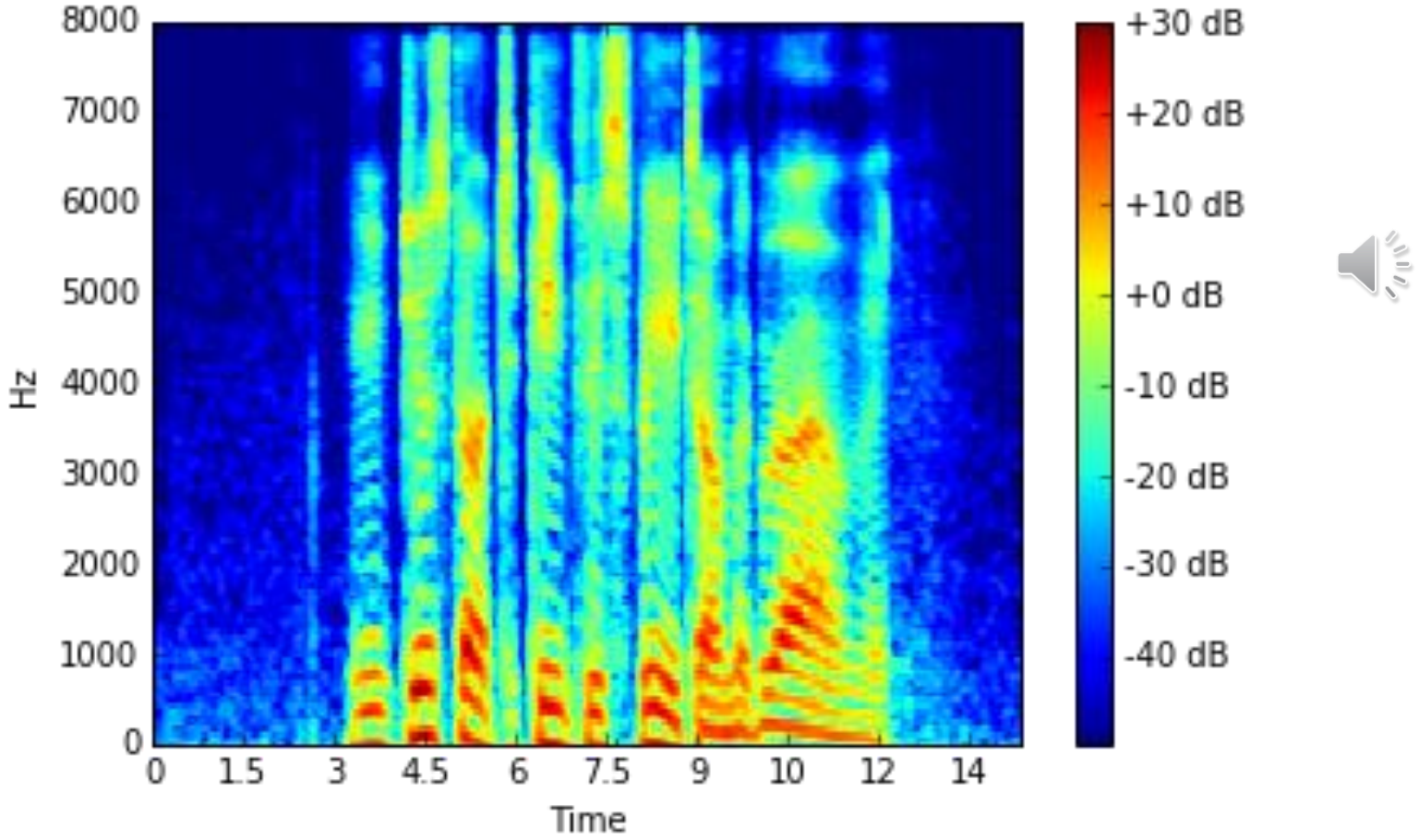
Example: End to end speech denoise

SNR: 2.5 dB
Café noise

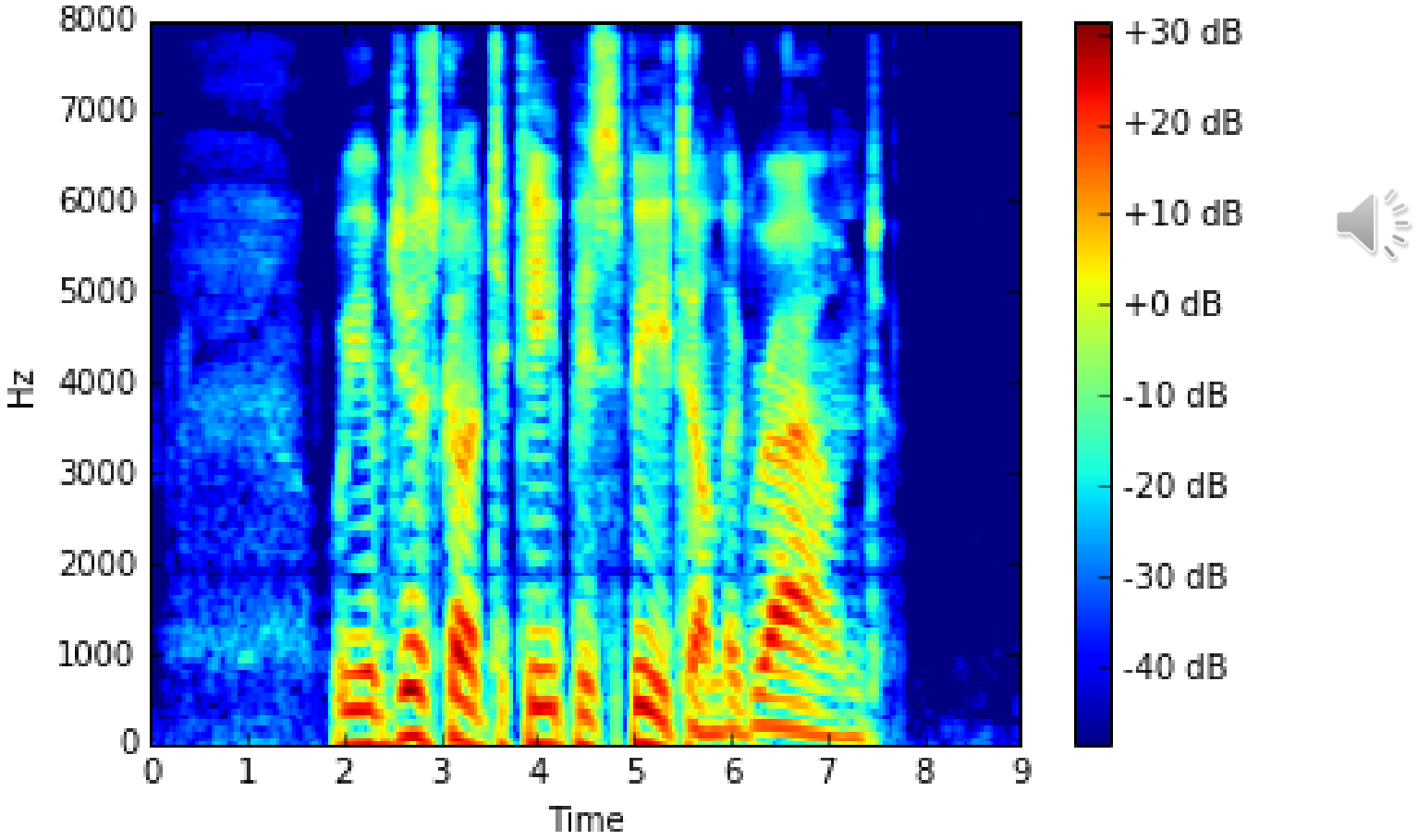


Example: Parametric resynthesis

SNR: 2.5 dB
Café noise



Reference: Clean speech



Motivation

- Standard speech enhancement → modify noisy recordings
  - Introduce distortions in speech
- Resynthesize clean speech from noisy mixture
- Use speech synthesis for speech enhancement
  - Vocoders to synthesize from acoustic features
  - Easier task than synthesis![1]

1. Maiti, S. and Mandel, M.I., 2019, May. "*Speech denoising by parametric resynthesis*", *ICASSP, 2019*

Parametric Resynthesis (PR)

- Speech enhancement using vocoders
  - Predict acoustic features from noisy speech
  - Synthesize speech from acoustic features
- High quality speech generation → neural vocoders[2]

Neural vocoders work on unseen speakers?

2. Maiti, S. and Mandel, M.I., 2019, October. "P*arametric resynthesis with neural vocoders*", *WASPAA 2019*

Neural Vocoders

- Models speech in time domain directly

WaveNet[3]	WaveGlow[4]	LPCNet[5]

• Autoregressive
• High Quality
• Slower synthesis
	• Generates samples in parallel
	• Glow based model
	• Fast synthesis
		• Autoregressive
		• Faster synthesis → written in C
		• Hybrid model
		  ○ Models vocal response with LPC coefficients
		  ○ Predicts excitation → simpler task

WaveNet: we use GPU accelerated nv-WaveNet – for faster synthesis

3. Oord, Aaron van den, et al. "*Wavenet: A generative model for raw audio*." , SSW, 2016.
4. Prenger, R, Rafael V, and Bryan C. "*Waveglow: A flow-based generative network for speech synthesis*.“, *ICASSP 2019*
5. Valin, J. M., & Skoglund, J. “LPCNet: Improving neural speech synthesis through linear prediction. ”, *ICASSP 2019*

Train neural vocoders with large number of speakers

- **Train**[6]**: 56** speakers
  - Voices from VCTK dataset
  - 28 male / 28 female
  - Accent: US and Scotland
- **Test**[6]**: 6** unseen speakers
  - 3 male / 3 female
  - Accent: England
- Sampling rate 16 kHZ
- **Objective quality metrics**
  - **CSIG, CBAK, COVL**[7]
  - 0-5 → higher is better

6. Valentini-Botinhao, C., 2017. Noisy speech database for training speech enhancement algorithms and TTS models.
7. Loizou, P.C., 2013. *Speech enhancement: theory and practice*. CRC press.

Speaker independence of neural vocoders

Scores averaged over 10 files/speaker

nv-Wavenet → low generation quality

Neural vocoders generalizes to unseen speakers

	CSIG	CBAK	COVL
1 Seen speaker			
WaveGlow	4.7	3.0	4.0
LPCNet	3.8	2.2	2.9
nv-WaveNet	3.3	2.1	2.5
6 Unseen speakers			
WaveGlow	4.6	2.8	3.9
LPCNet	4.0	2.4	3.1
nv-WaveNet	3.2	2.1	2.5

Effect of gender on neural vocoders

Scores averaged over 10 files/speaker

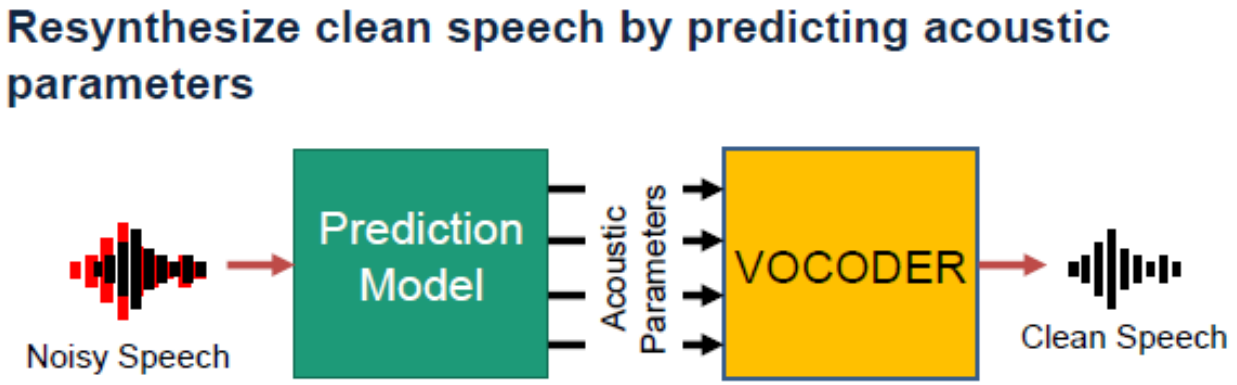
WaveGlow models female speakers slightly better

LPCNet, WaveNet no such difference

	CSIG	CBAK	COVL
Male			
WaveGlow	4.5	2.8	3.8
LPCNet	4.0	2.3	3.1
nv-WaveNet	3.2	2.1	2.5
Female			
WaveGlow	4.6	2.8	3.9
LPCNet	4.0	2.4	3.2
nv-WaveNet	3.3	2.0	2.5

Parametric Resynthesis (PR)

1. Predict “clean” acoustic features from noisy speech
2. Reconstruct speech from acoustic features
   - Acoustic features are different for different vocoders



Vocoder acoustic features

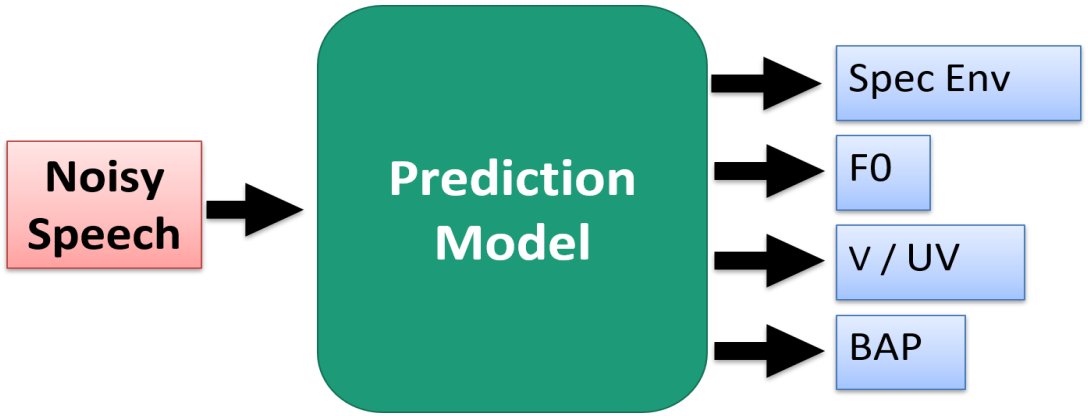
Vocoder	Acoustic features	Dim
WaveNet	Mel-spectrogram	80
WaveGlow	Mel-spectrogram	80
LPCNet	BFCC, F0 period, F0 correlation	20
WORLD	Spectral envelope, aperiodicity, F0, v/uv	63

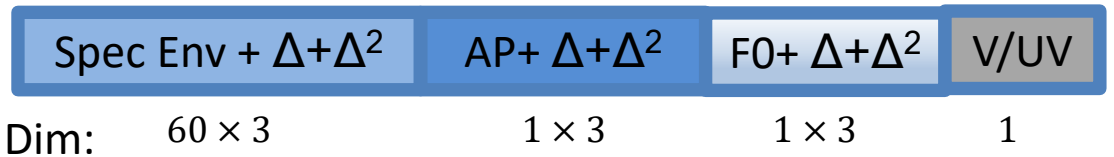
Neural vocoders →

WaveNet, WaveGlow, LPCNet

Prediction model

- Predicts acoustic features (X) at a fixed frame rate
  - **Input:** noisy mel-spectrogram (Y)
  - **Loss:** $\mathrm{MSE} := |X - X'|^2$
  - X' → predicted features





PR-WORLD

Training of Parametric resynthesis

- **Noisy trainset:** 56 speaker set
  - 8 noises from DEMAND
  - 2 artificial noises
  - SNR range: 15 – 0 dB
- **Test:** 2 unseen speakers
  - 8 unseen noises from DEMAND
  - 824 files
  - 4 SNR level: 17.5 dB, 12.5 dB, 7.5 dB, 2.5 dB
- Comparison models:
  - SEGAN[8], Wave-U-Net[9], Wavenet-denoise[10]
  - Oracle Wiener mask
    - Has access to clean speech

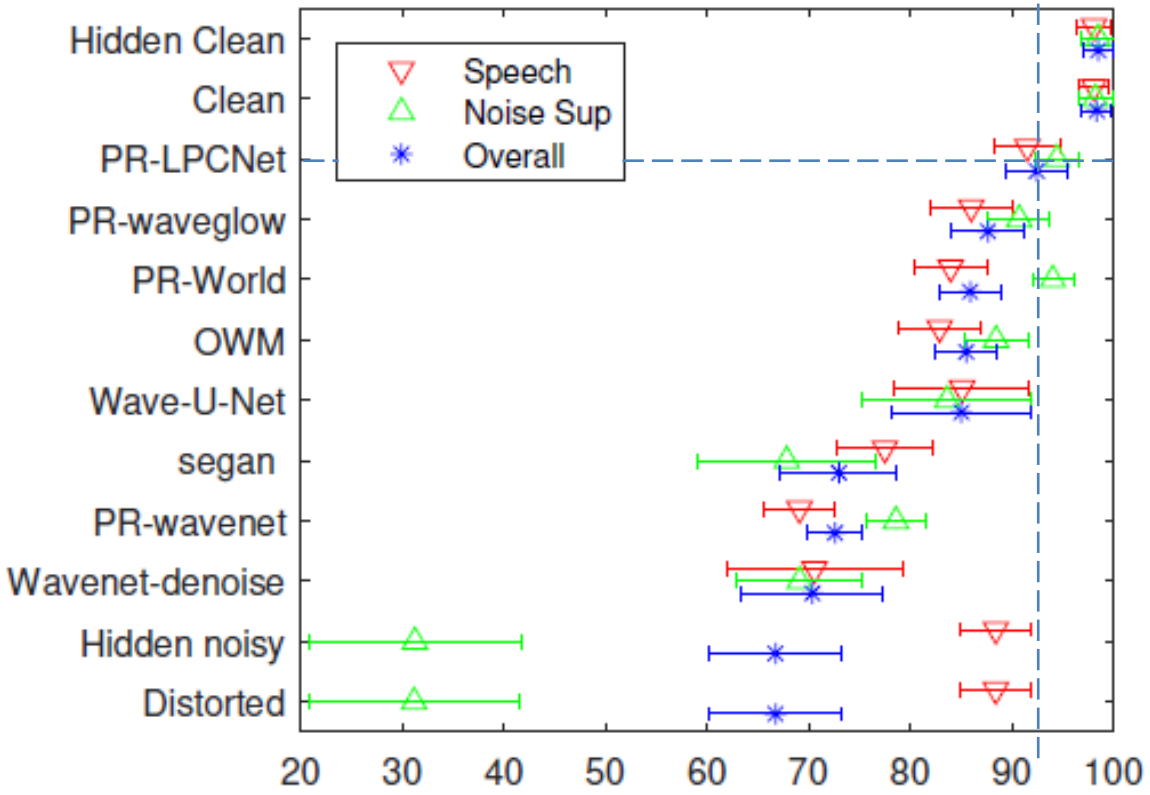
8. Pascual, S., Bonafonte, A. and Serra, J., "SEGAN: Speech enhancement generative adversarial network". *arXiv, 2017.*
9. Macartney, C. and Weyde, T., "Improved speech enhancement with the wave-u-net". *arXiv, 2018.*
10. Rethage, D., Pons, J. and Serra, X. "A wavenet for speech denoising",*ICASSP, 2018*.

Objective metrics for speech enhancement

PR-WaveGlow performs best in CSIG and CBAK

	CSIG	CBAK	COVL	STOI
Oracle Wiener	4.3	3.8	3.8	0.98
PR-WaveGlow	3.8	2.4	3.1	0.91
PR-LPCNet (noisy F0)	3.5	2.1	2.7	0.88
PR-LPCNet	3.1	1.8	2.2	0.88
Wave-U-Net	3.5	3.2	3.0	
SEGAN	3.5	2.9	2.8	

Subjective Quality



- MUSHRA listening test
- Number of files: 12
- SNR: 12.5 dB to 2.5 dB

PR-LPCNet outperforms all systems!

Objective metrics on 12 listening test files

	CSIG	CBAK	COVL
OWM	4.3	3.8	3.9
PR-WaveGlow	**3.8**	**2.4**	**3.0**
PR-World	3.1	1.9	2.2
PR-LPCNet	**3.0**	**1.8**	**2.2**

Subjective quality scores does not match objective scores

LPCNet scores 0.8 lower than WaveGlow!

In a nutshell

- Neural vocoders
  - Speaker Independent when trained on large number of speakers
  - All 3 vocoders were able to generalize to unseen speakers
- Speech enhancement
  - PR-LPCNet
    - Outperforms Oracle Wiener mask in subjective quality scores
  - PR-WaveGlow
    - Higher objective metrics than LPCNet

Thank You