Exciting Estimated Clean Spectra Speech Resynthesis

Sreyas Srimath Tirumala & Micheal I Mandel
srimath-tirumala.1@osu.edu, mim@sci.brooklyn.cuny.edu

The Ohio State University

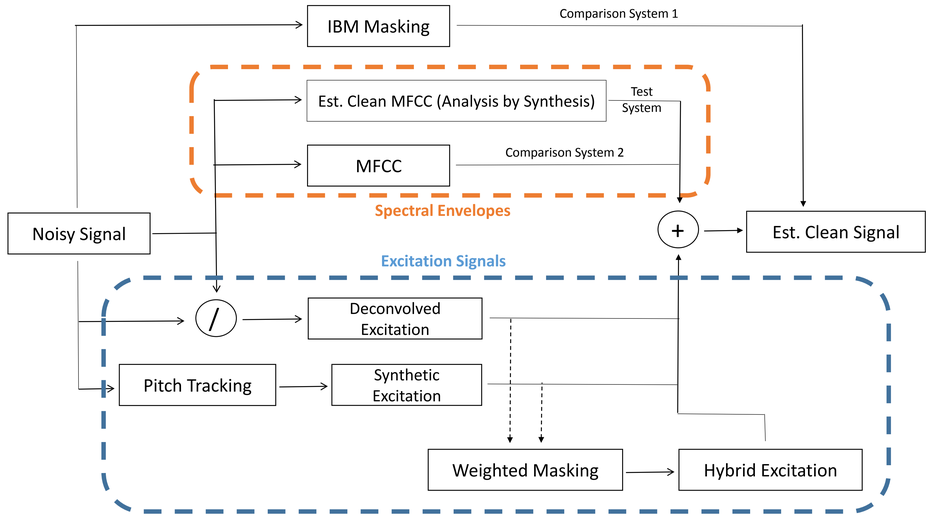
Problem

- Does the estimated clean spectral envelope from Mandel and Narayanan, 2014 (AbS) sound cleaner than the noisy or spectral masked speech?
- Can a good enough estimate of the excitation signal be generated to resynthesize the signal from the spectral envelope with minimal errors?
- Premise:
  - Estimate of the clean spectral envelope and best clean estimate of the excitation should result in better signal than masked noisy signals
  - Spectral Masking can damage speech by allowing noise to bleed in regions where both speech and noise are present
- Target:
  - The estimated clean speech should have better noise suppression than spectral masked speech

Approach

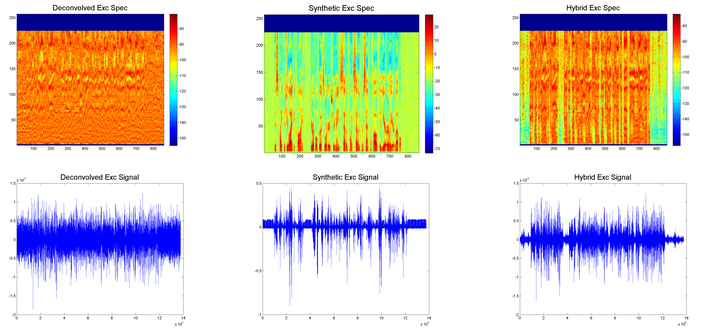
- Treat clean speech resynthesis as finding
  - Estimate of the spectral envelope from AbS
  - Estimate of excitation
- Excitation signal estimated using the following methods:
  - Deconvolution of the signal using the spectral estimate
  - Synthetic excitation generated using the signal pitch
  - Hybrid excitation: weighted combination of the above

Resynthesis System Flow



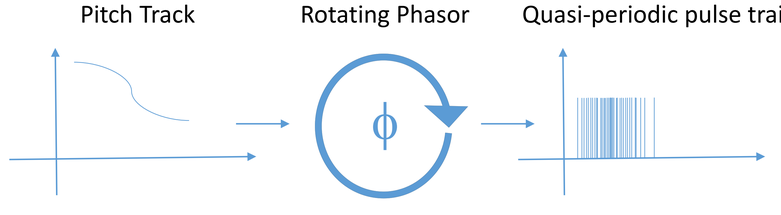
Example: Excitation Signals

- Deconvolved excitation has noise visible in unvoiced regions
- Synthetic excitation has less energy in the voiced regions
- Hybrid excitation reduces noise and preserves voiced structure.



Excitation Models

- Excitation from deconvolution: Deconvolving the spectrum by the estimated clean speech spectral features. This approach can lead to noise bleeding since the observation is noisy.
- Synthetic Excitation from pitch:
  - Quasi-periodic pulse train can be generated by integrating the time-varying pitch and exciting impulses based on the rotation of the phasor
- Hybrid Excitation: Formed by interpolating them both in the time-frequency domain.



Results

Spec	Exc	Quality Speech	Quality Noise	Quality Overall
Clean		91.93	91.64	92.46
Noisy		85.91	32.96	61.56
DM		**69.13**	59.51	65.43
Noisy	Hyb	64.43	76.06	68.06
Est	Hyb	58.88	**87.26**	**68.97**

Legend: Results compiled from the MUSHRA listening test taken by 6 listeners for 15 versions of 5 utterances with 6 variations of noise. The higher rating corresponds to better quality / noise suppression. Spec refers to the spectrum (or features) used and Exc refers to the excitation signal used.

- For noise suppression, signals resynthesized using the spectral features from AbS were rated highest
- Better estimate of the excitation can improve resynthesis
- The test signal (using estimated MFCC and the hybrid excitation) is second only to the clean signal in terms of noise suppression
- Statistical test of significance proved ($p < 10^{-6}$) that the hybrid excitation with estimated spectral features suppresses noise more than it damages speech.

Future Directions

- Results show that a better estimate of the excitation can significantly improve resynthesis quality
- Better estimate of the phase
  - Learning invertible phase features using a DNN/RNN
  - Joint learning of the phase and amplitude for a noisy signal using complex DNN/RNN
- Adding the estimation of the excitation signal into the analysis by synthesis framework

References

Mandel, M. I. and A. Narayanan (2014). "Analysis-by-synthesis feature estimation for robust automatic speech recognition using spectral masks". In: Proceedings of the IEEE ICASSP.