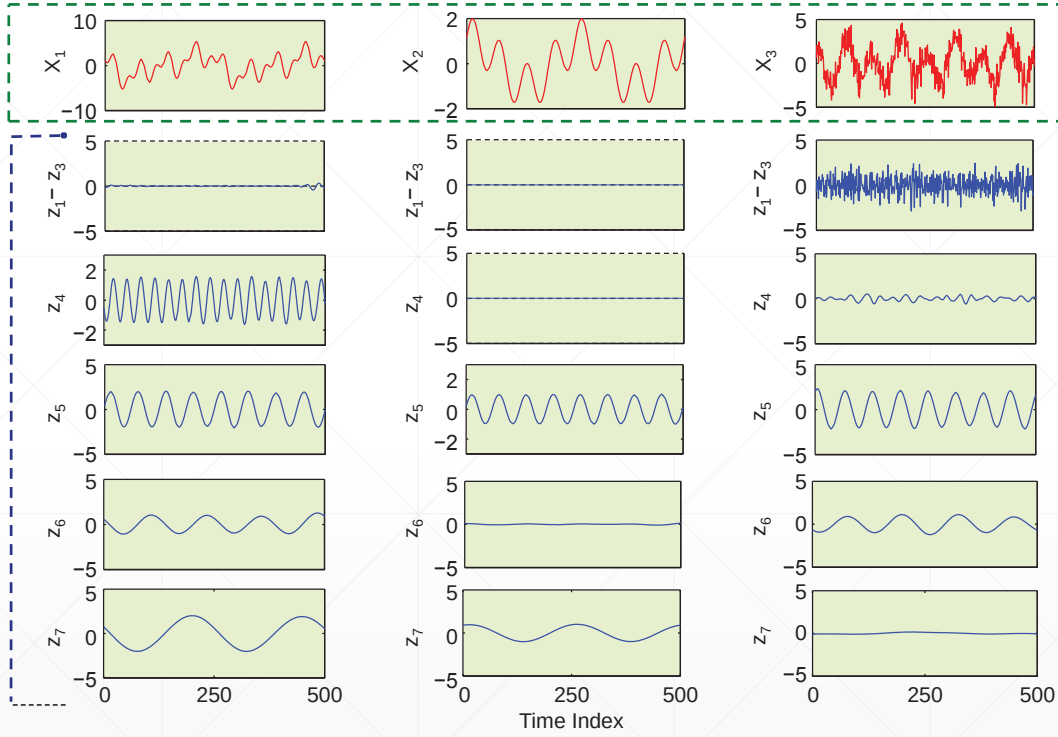
Robust Speech Analysis Based on Source-Filter Model Using Multivariate Empirical Mode Decomposition in Noisy Environments

Surasak Boonkla[1,2], Masashi Unoki[1], and Stanislav S. Makhanov[2]

[1]School of Information Science, Japan Advanced Institute of Science and Technology, Japan
[2]Sirindhorn International Institute of Technology, Thammasat University, Thailand

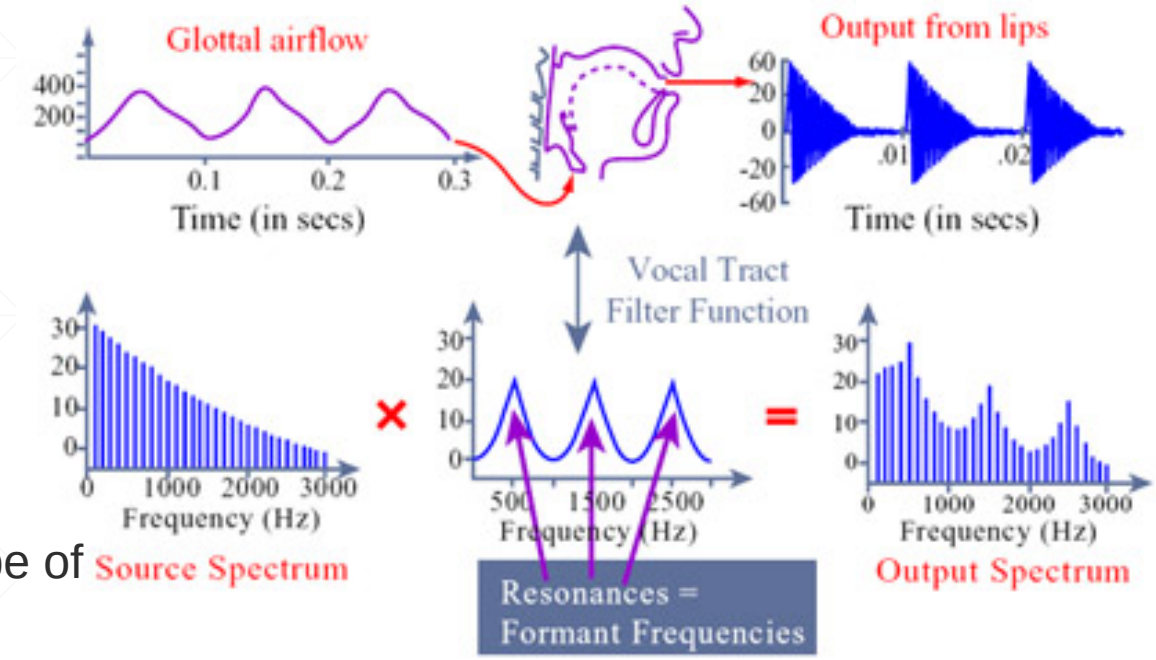
Outline

- Introduction
- Multivariate Empirical Mode Decomposition (MEMD)
  - “COMMON MODE alignment”
- Proposed Method
- Evaluation and Result
- Discussion
- Conclusion



Introduction

- Speech analysis tries to capture
  - Glottal-source: fundamental frequency (F0)
  - Vocal-tract: Resonances (F1, F2, F3, …), Shape of spectral envelope
- Applications
  - Voice activity detection (VAD)
  - Automatic speech recognition (ASR), etc.
- Existing methods
  - Linear prediction (LPC), Cepstrum (CEP)
  - ARMA, AbS, STRAIGHT



Source-filter model

$$s(t) = e(t) * v(t)$$

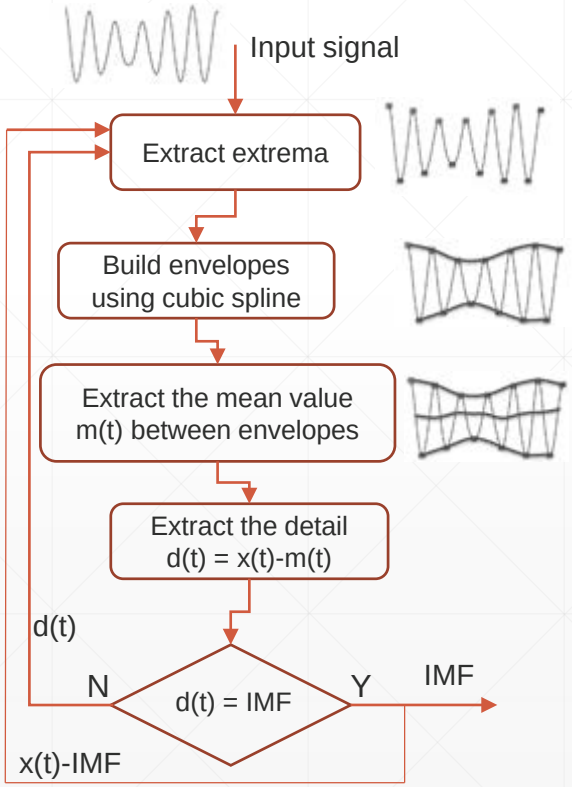
Problem: robustness in noisy environments (low SNR).

Aim: propose a robust speech analysis method.

Multivariate Empirical Mode Decomposition (MEMD)

$$x(t) = \sum_{i=1}^{K} z_i(t) + r(t)$$

Intrinsic Mode Function (IMF), $z_i(t)$

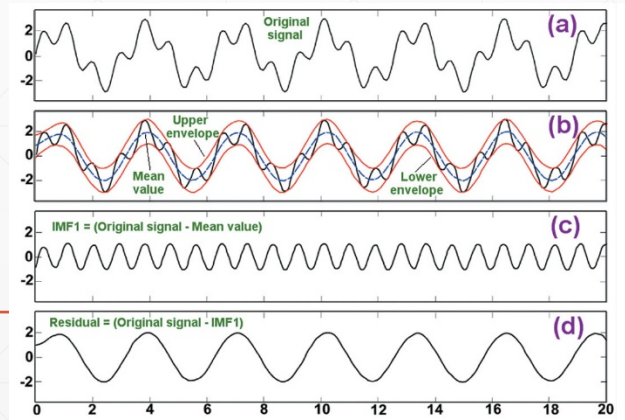


Multivariate EMD [2]

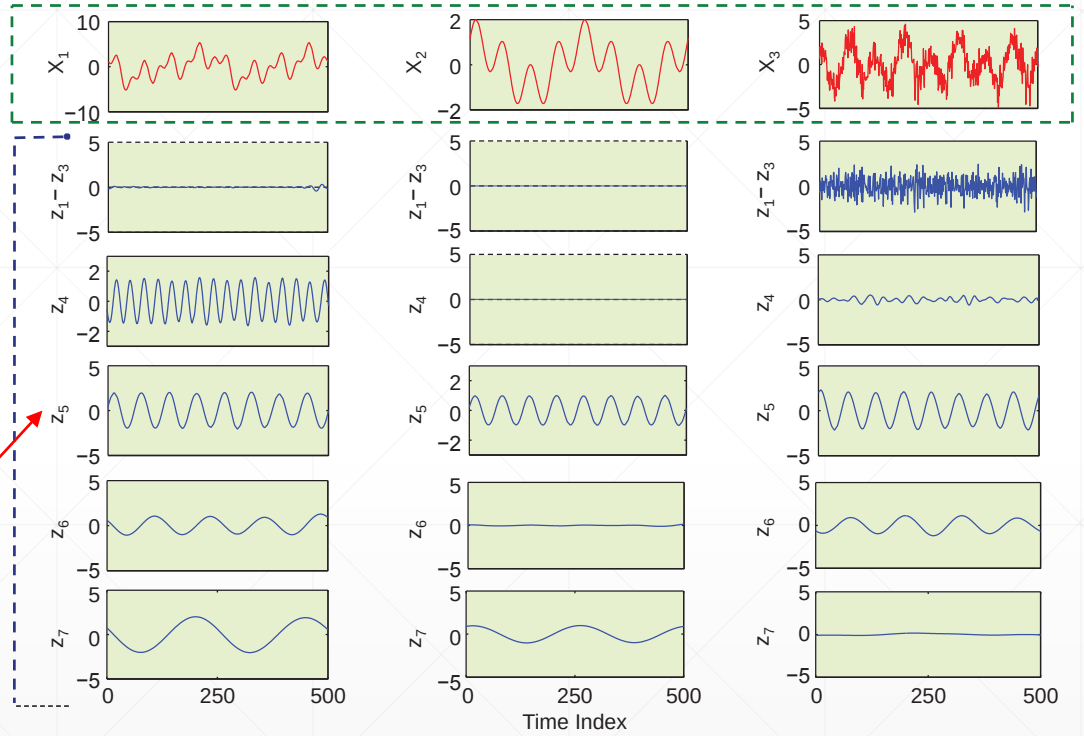
X = [$x_1(t)$, $x_2(t)$, $x_3(t)$]

$x_1(t)$ = [8 Hz + 16 Hz],
$x_2(t)$ = [8 Hz + 4 Hz],
$x_3(t)$ = [8 Hz + 2 Hz + noise].

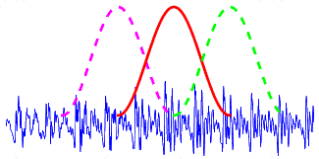
Common Mode

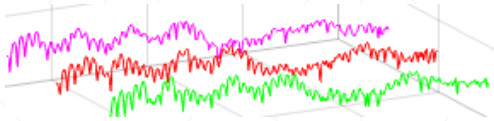


Hansan [3]



Common Mode (1)





Common Mode

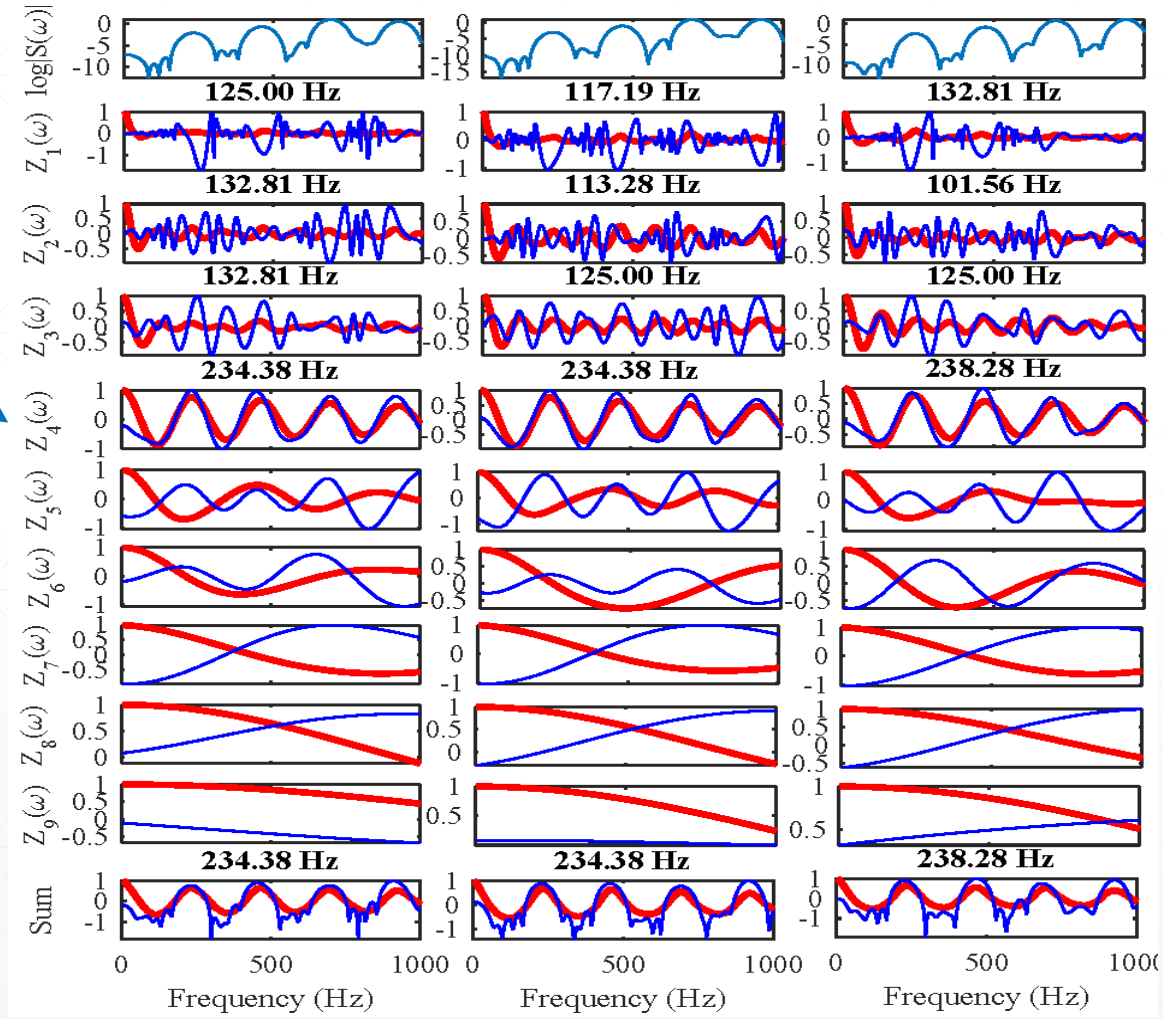
Glottal-source waveform

Impulse response of vocal-tract filter

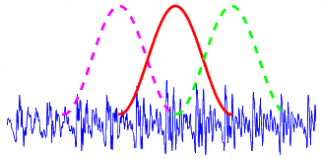
$$s(t) = e(t) * v(t)$$
$$S(\omega) = E(\omega)V(\omega)$$
$$\log|S(\omega)| = \log|E(\omega)| + \log|V(\omega)|$$

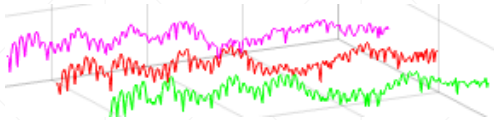
$$\log|S(\omega)| = \underbrace{\sum_{i=1}^{M} z_i(\omega)}_{\text{Source}} + \underbrace{\sum_{i=M+1}^{K} z_i(\omega)}_{\text{Filter}}$$

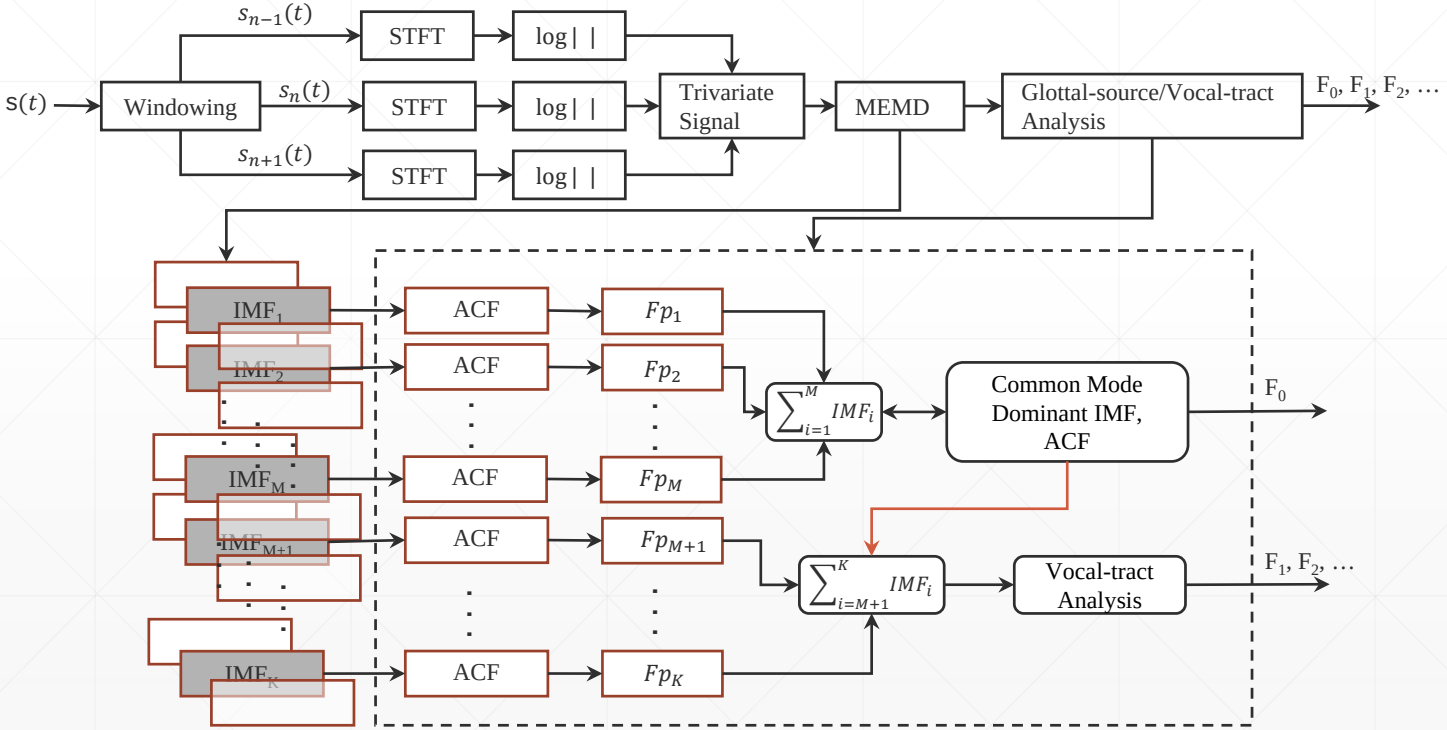
IMF, ACF



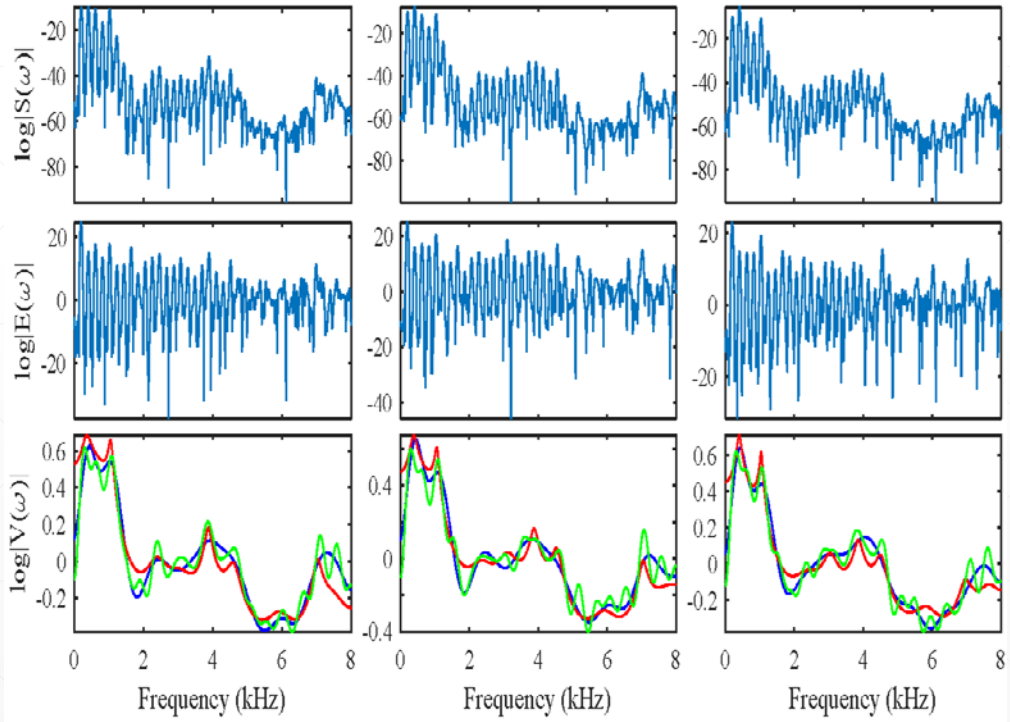
Proposed Method (1)





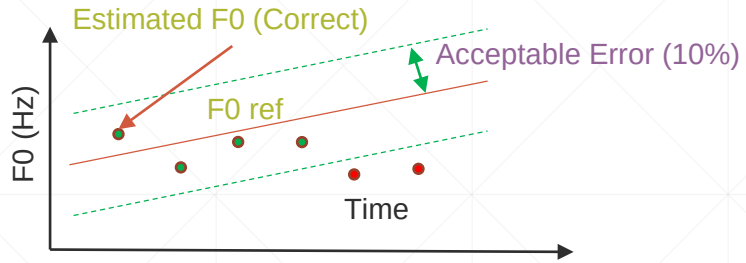


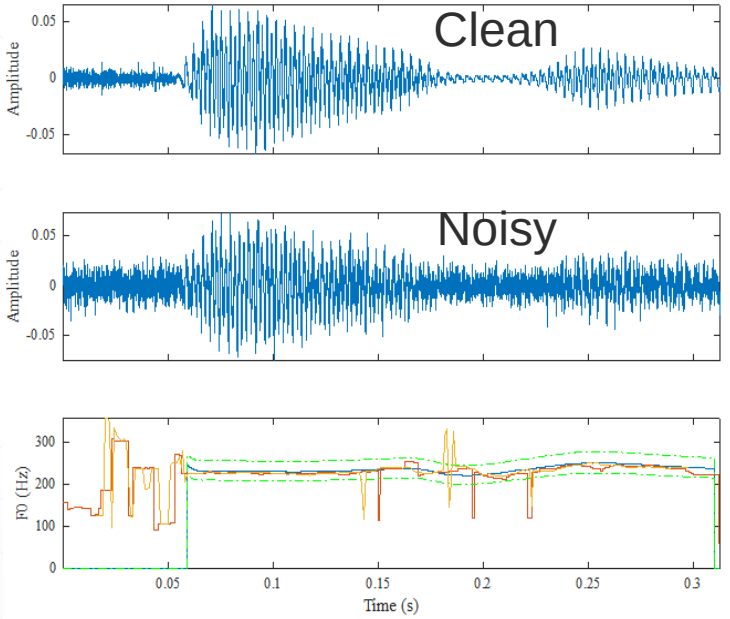
$$\log|S(\omega)| = \underbrace{\sum_{i=1}^{M} z_i(\omega)}_{\text{Source}} + \underbrace{\sum_{i=M+1}^{K} z_i(\omega)}_{\text{Filter}}$$

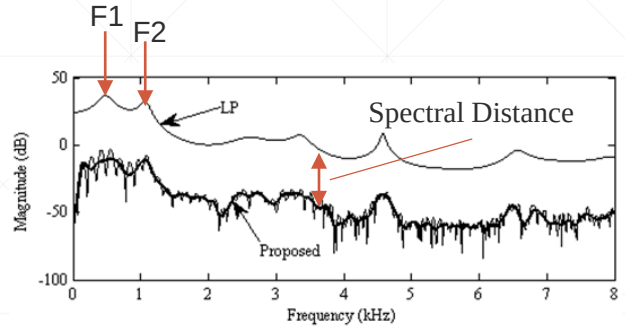


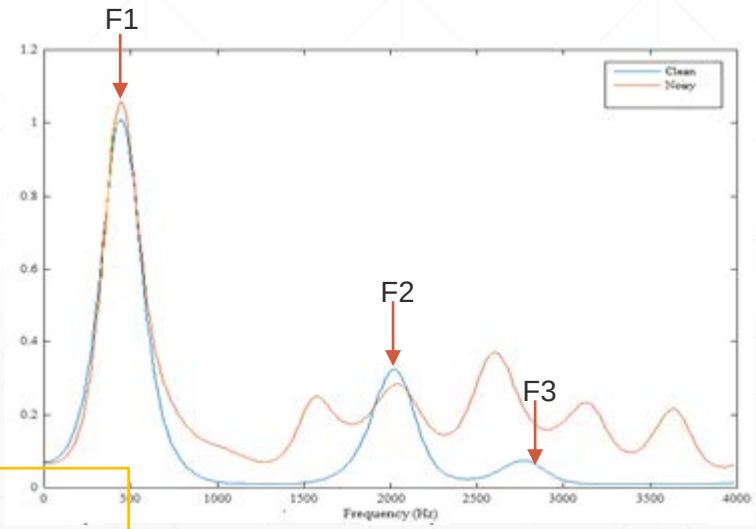
Evaluations

TEMPO for F0 ref.









■ Glottal-Source
- F0 estimation

$$\text{Correct rate}[4] = \frac{No.Correct}{No.All} \times 100$$

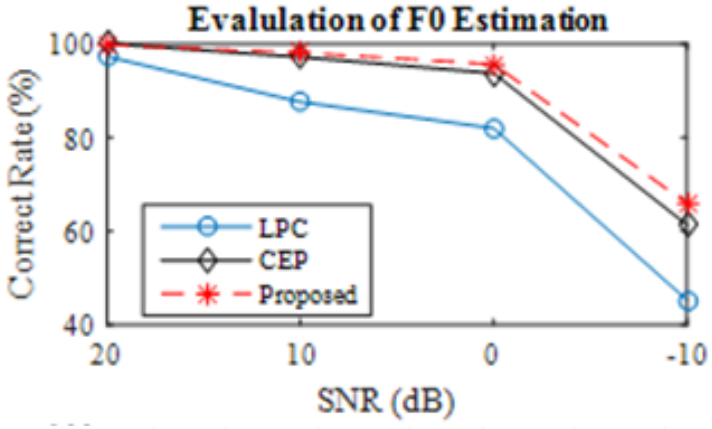
Cepstrum, MEMD, TEMPO, Err

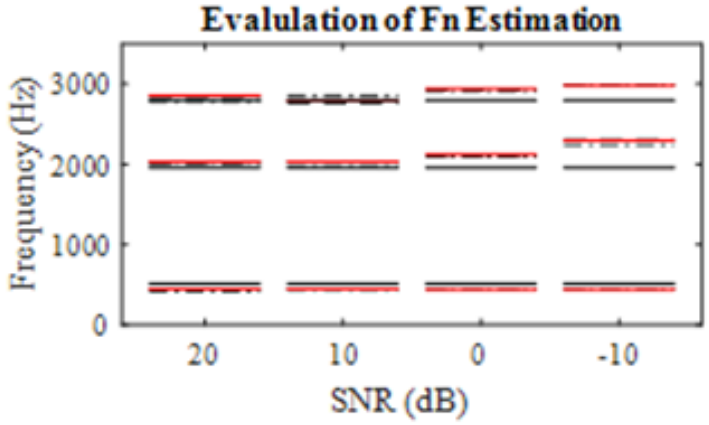
■ Vocal-tract
- Formants (F1, F2, F3): Plots of formants compared with the correct ones
- Shape: Correlation Coefficient and Euclidean Spectral Distance

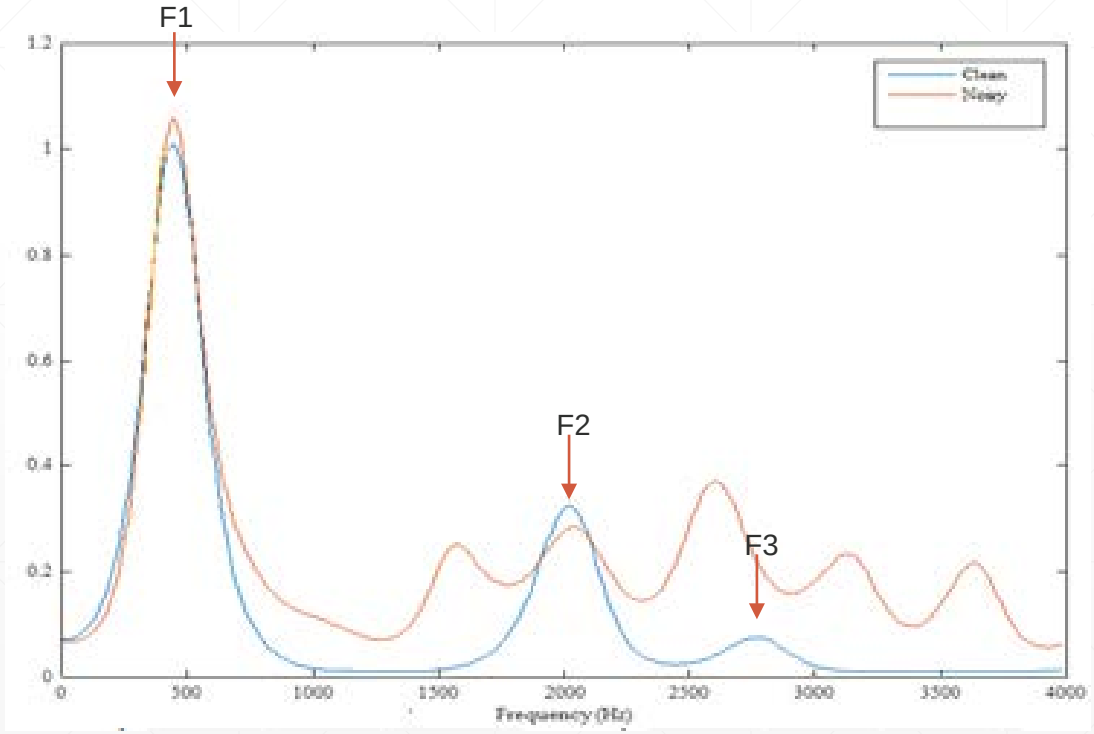
Stimuli: voiced speech signal vowel /ey/ from 10 persons (TIMIT database), SNR: 20, 10, 0, and -10 dB

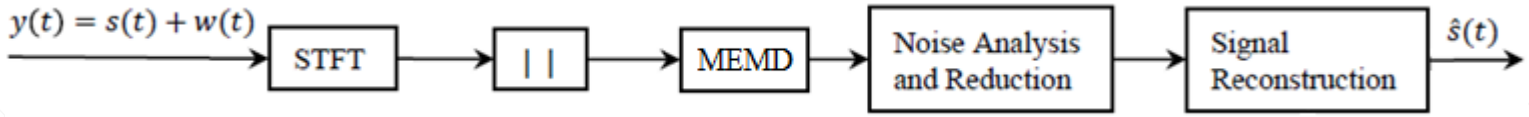
Result (1)

- F0 estimation was robust.
- Formant estimation was not robust when SNR were 0 and -10 dB.
- Spectral envelope was bad.
- Formant estimation and the shape of spectral envelope was required to be improved.





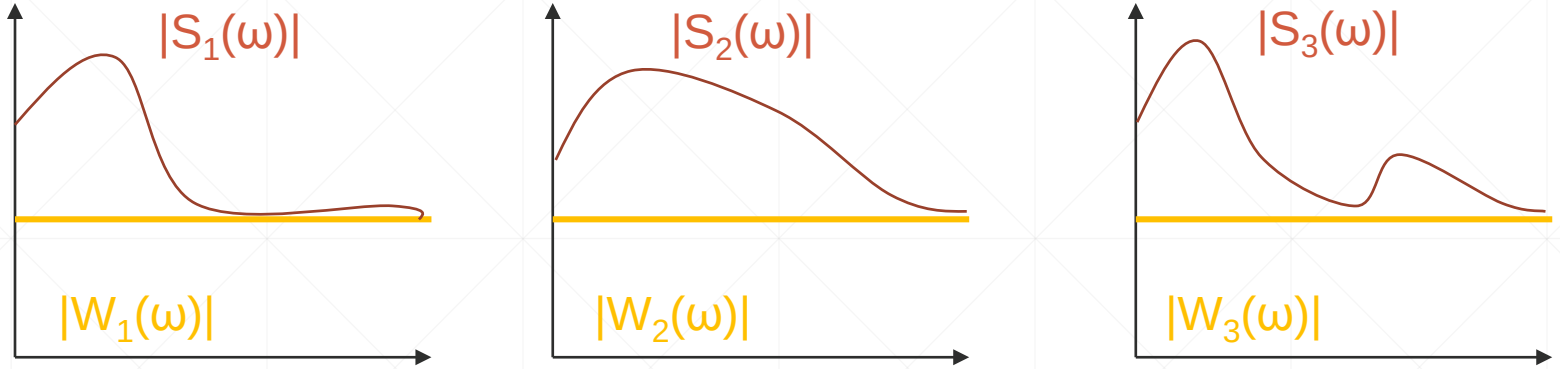




Common Mode (2)

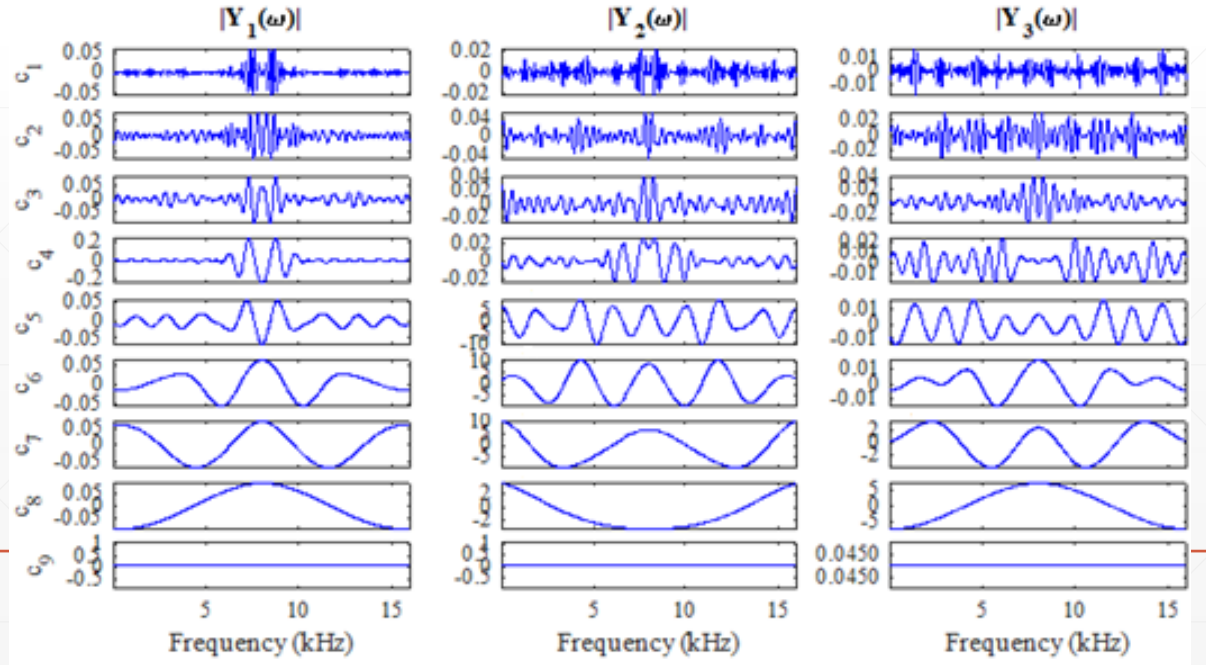
$$|Y(\omega)| = |S(\omega)| + |W(\omega)|$$

$$|Y(\omega)| = \underbrace{\sum_{i=1}^{L} z_i(\omega)}_{\text{Speech}} + \underbrace{\sum_{i=L+1}^{K} z_i(\omega)}_{\text{Noise}}$$

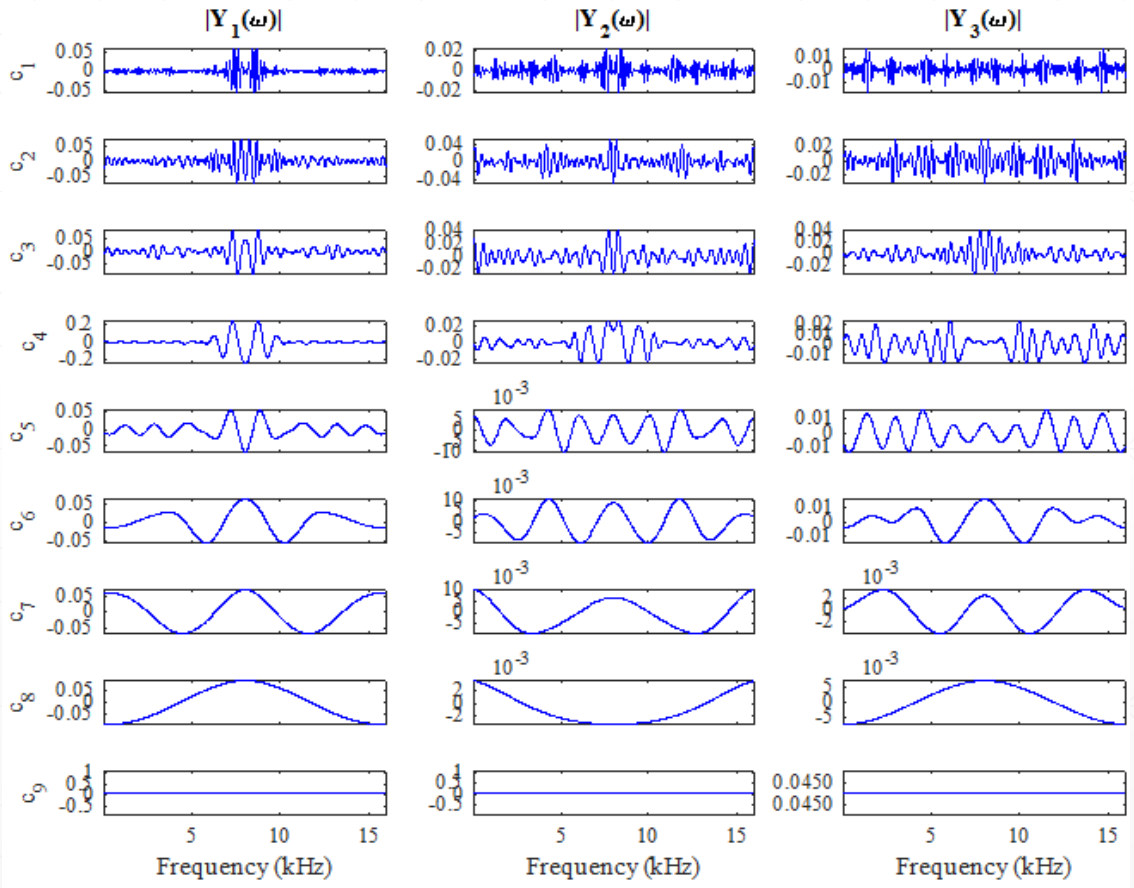


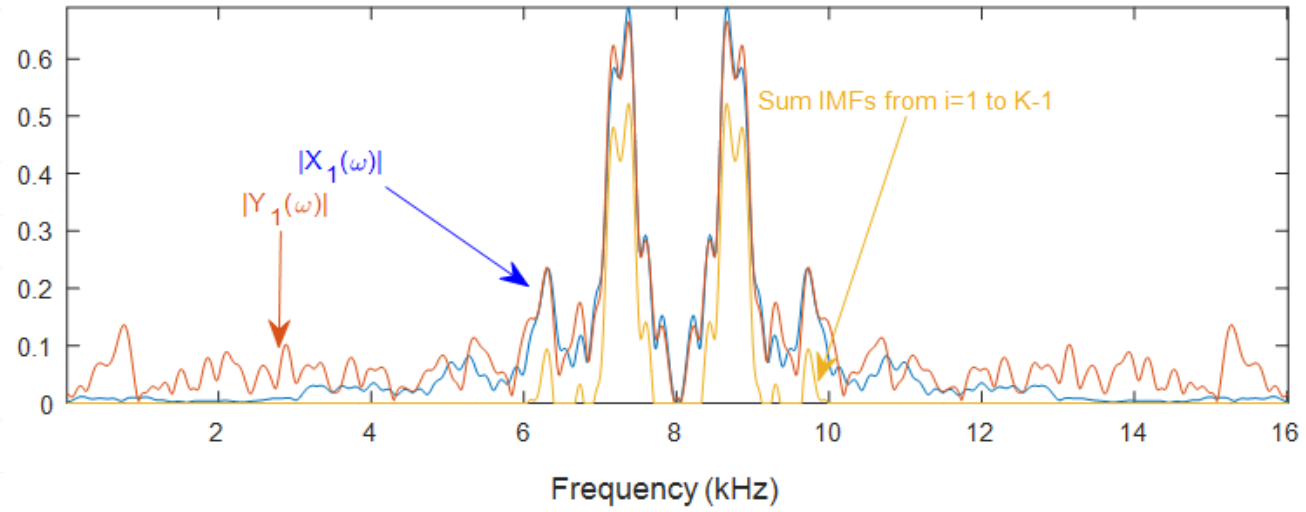
Assumptions

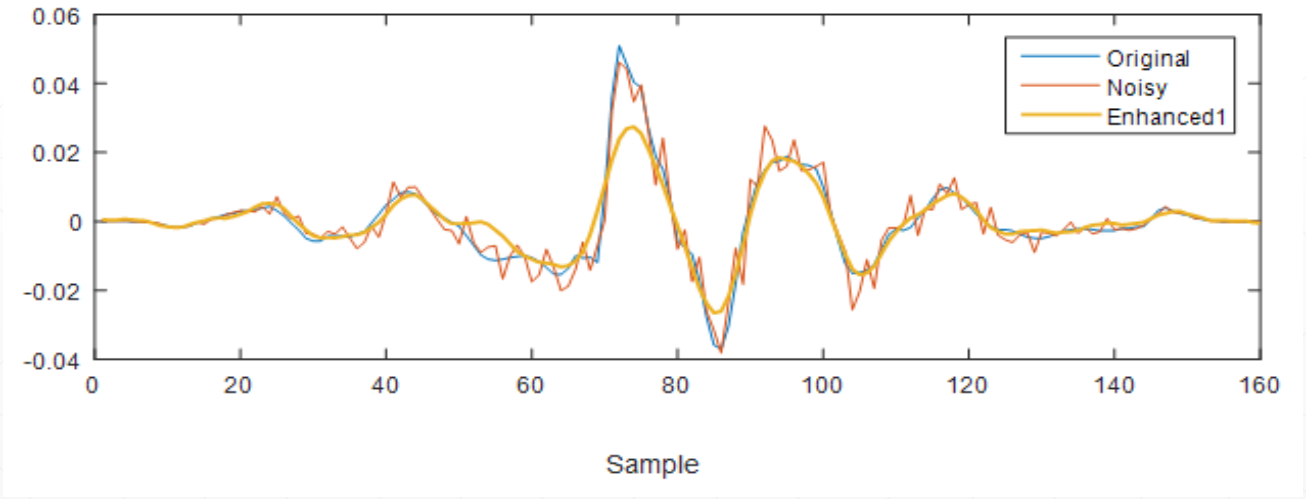
- Noise is stationary.
- Speech is non-stationary.



Common Mode (4)







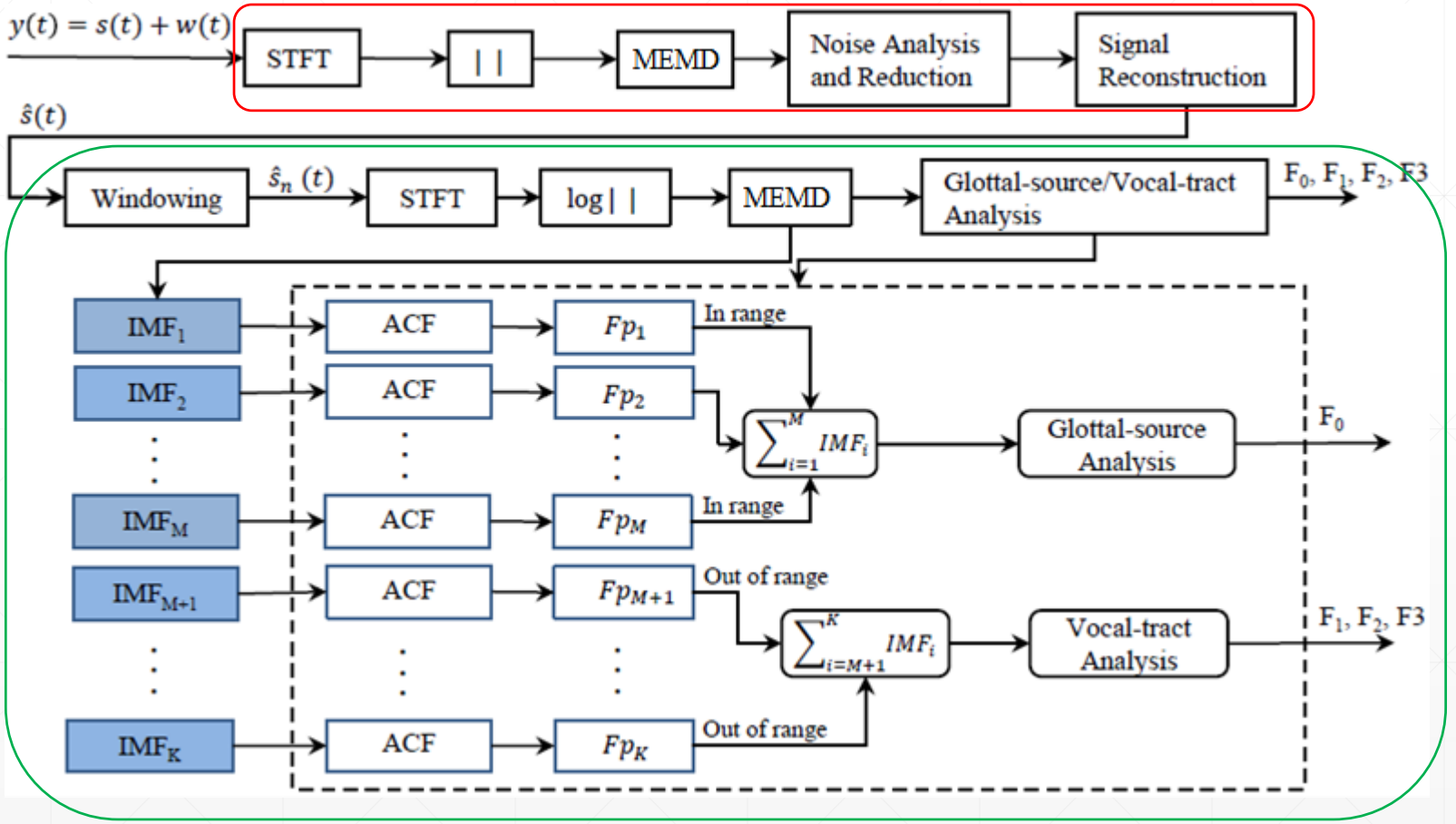
Proposed Method (2)

5-10 ms of winLen, No overlap

MEMD-Based Noise Analysis and Reduction

Two-stage speech analysis

- 1st stage: MEMD-based noise analysis and reduction
- 2nd stage: MEMD-based speech analysis

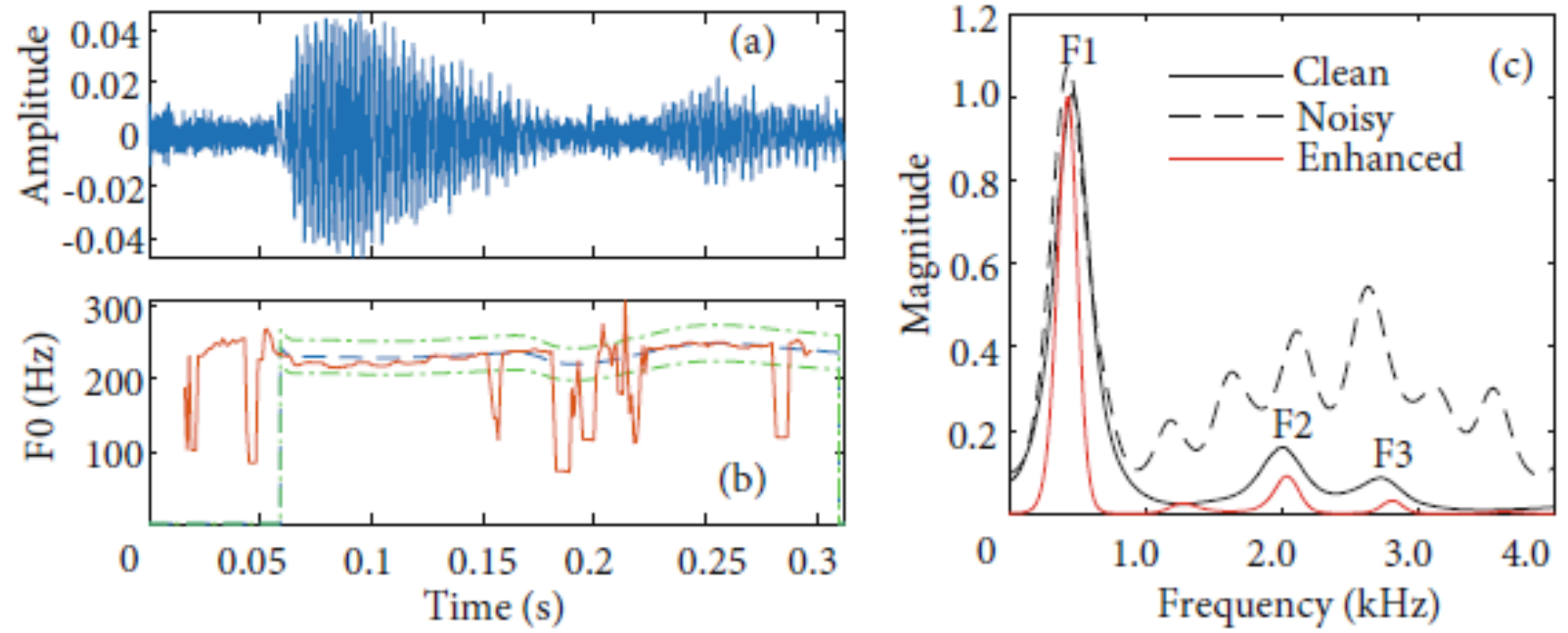


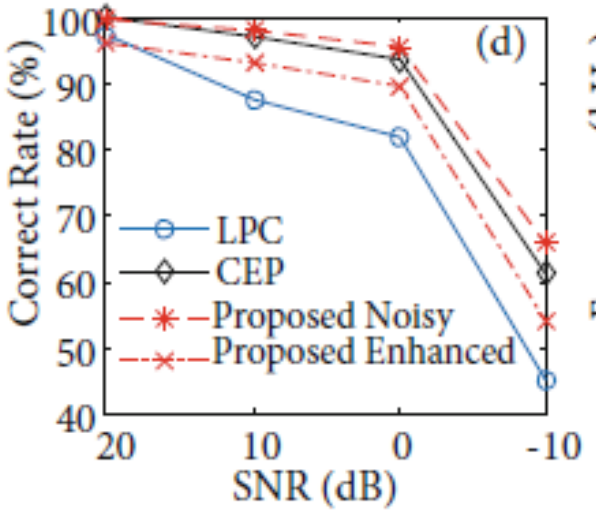
MEMD-Based Speech Analysis

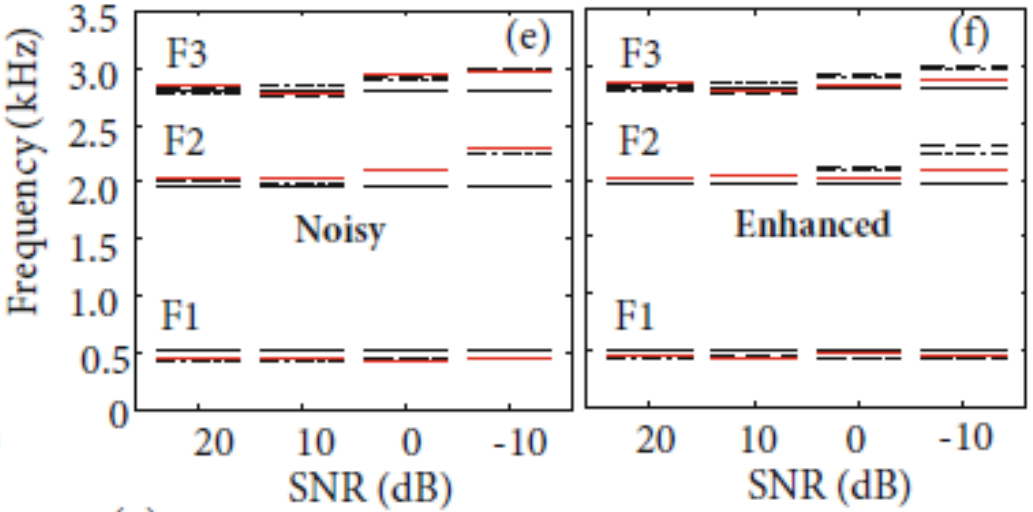
30 ms of winLen, 50% overlap

Results (2)

- F0 estimation outperforms LPC and CEP before noise reduction.
- Formants estimation was improved after noise reduction.
- Shape of spectral envelope was improved.
  - Correlation coefficient increases.
  - Spectral distance decreases (0 and -10 dB).







(g)

	SNR (dB)							
	20		10		0		-10	
	Dist	Corr	Dist	Corr	Dist	Corr	Dist	Corr
Noisy	2.40	0.81	2.33	0.84	4.33	0.83	10.31	0.57
Enhanced	*3.34*	**0.90**	*3.35*	**0.91**	**2.97**	**0.87**	**3.74**	**0.75**

Discussion (1)

- Before noise reduction (2nd stage)
  - F0 estimation was better than LPC and CEP. ✓
  - Formant estimation was not robust when SNR were 0 and -10 dB. ✗
  - Spectral envelope was bad. ✗
- After noise reduction (1st and 2nd stages)
  - Correct rate of F0 estimation was reduced. ✗
  - Formant estimation was improved. ✓
  - The shape of spectral envelope was improved. ✓
- Combining two stages leads to robust speech analysis

Conclusion

- Proposed robust speech analysis method based on source-filter model using MEMD.
- Automatically decomposed noise as the common mode.
- Automatically separate source and filter using the common mode.
- The proposed method could be robust in noisy environments.
- Future work: pink noise, babble noise, reverberation

References

[1] N. E. Huang, "**The Empirical Mode Decomposition and the Hilbert Spectrum for Non-Linear and Non-stationary Time Series Analysis**," Proc. the Royal Society: Math, Physi., and Eng. Sci., A454, 903-995, 1998.

[2] D. P. Mandic, N. U. Rehman, Wu Zhaohua, and N. E. Huang, "**Empirical Mode Decomposition-Based Time-Frequency Analysis of Multivariate Signals: The Power of Adaptive Data Analysis,** " IEEE Signal Processing Magazine, Vol. 30 , No. 6, pp. 74 - 86 , Nov. 2013.

[3] Hassan H. Hassan and John W. Peirce, "**Empirical Mode Decomposition (EMD) of potential field data: airborne gravity data as an example**, " Canadian Society of Exploration and Geophysicists **(**CSEG), VOL. 33 No. 01, Jan 2008.

[4] S. Boonkla, M. Unoki, S. S. Makhanov, and C. Wutiwiwatchai, "**Speech analysis method based on source-filter model using multivariate empirical mode decomposition in log-spectrum domain**," IEEE International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 555-559, Sept. 2014.

[5] M. Unoki, T. Hosorogiya, and Y. Ishimoto, "**Comparative Evaluations of Robust and Accurate F0 Estimates in Reverberant Environments**," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4569-4572, Mar. 2008.