Speech Analysis Based on Source-Filter Model Using Multivariate Empirical Mode Decomposition

S. Boonkla[1,2], M. Unoki[1], Makhanov S. S.[2], and C. Wuthiwiwatchai[3]

[1]School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Japan
[2]School of Information, Communication and Computer Technologies (ICT), Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU), Thailand

[3]National Electronics and Computer Technology Center (NECTEC), Thailand

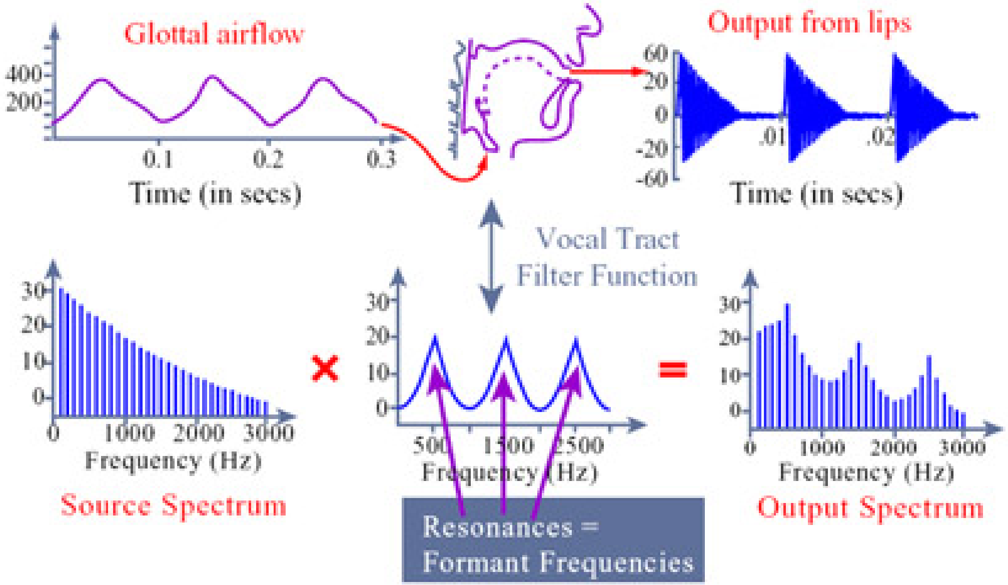
Motivation & Aim

- Linear Prediction (LP) separates glottal-source and vocal-tract filter based on sampling rate.
- Cepstrum separates glottal-source and vocal-tract filter using liftering the cut-off quefrency of which depends on gender.
- MEMD can automatically separate glottal-source and vocal-tract filter.

Introduction

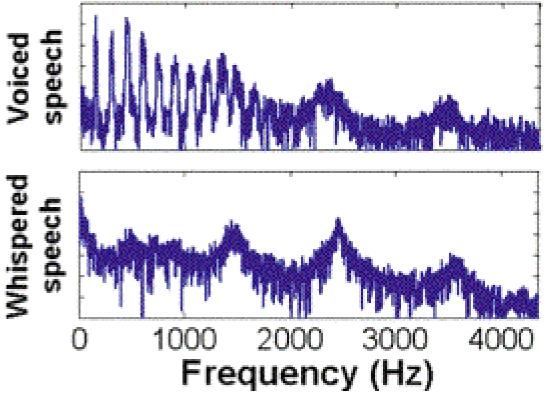
- Speech analysis is important for several applications such as automatic speech recognition systems, speech analysis/synthesis, hearing aids, etc.
- Existing speech analysis method are still weak in real environments.
- Improving the existing methods and finding a new one are important.

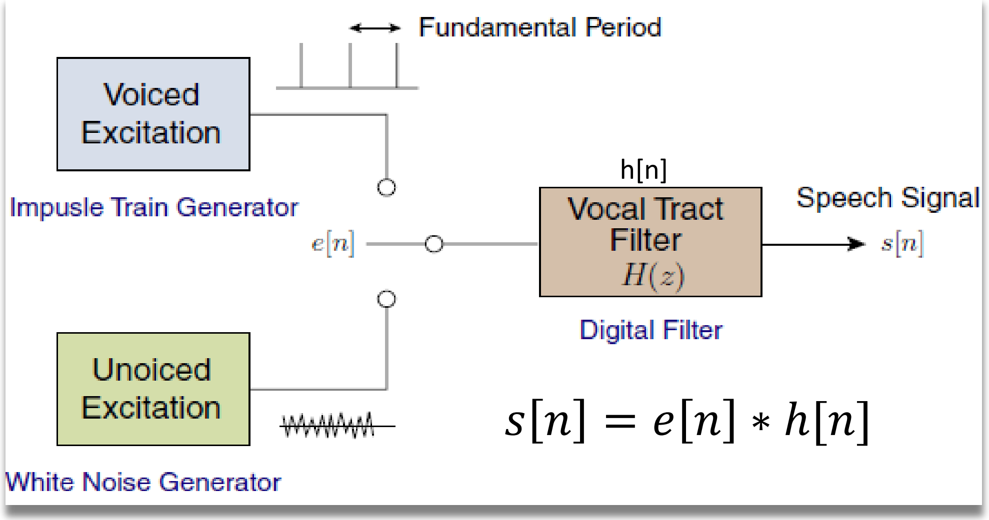
Source-Filter Model



mit.tsu.edu.ph/OcwWeb/Linguistics-and-Philosophy/24-963Fall-2005/CourseHome/index.htm

newt.phys.unsw.edu.au/jw/voice.html#vowels





- ❑ A speech signal, s[n], is resulted from convolution of glottal-source signal, e[n], and a vocal-tract filter, h[n].
- ❑ Glottal-source has two types which are voiced and unvoiced excitation. We consider voiced excitation here.

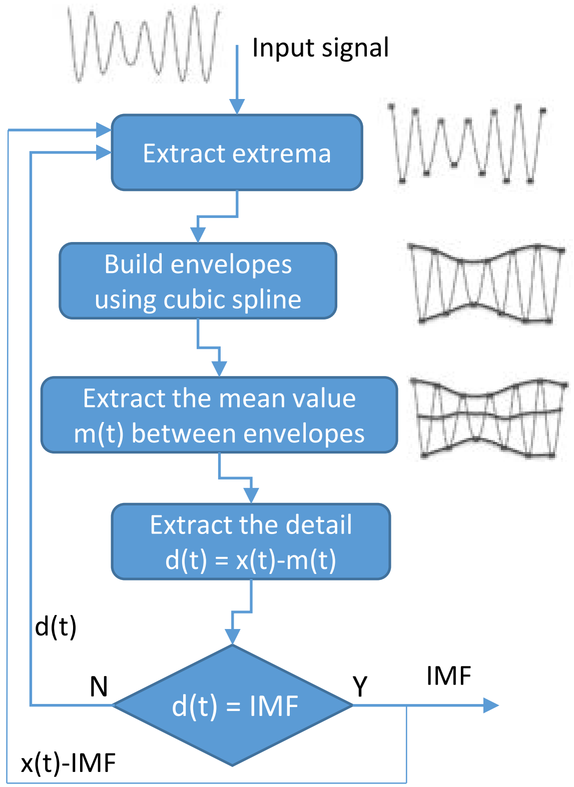
$$s[n] = e[n] * h[n] \xrightarrow{\text{DTF}} S[k] = E[k]H[k] \xrightarrow{\quad | \;\; | \quad} |S[k]| = |E[k]||H[k]| \xrightarrow{\text{Log}} \log|S[k]| = \log|E[k]| + \log|H[k]|$$

Multivariate Empirical Mode Decomposition

Empirical Mode Decomposition (EMD) [1]

$$x(t) = \sum_{i=1}^{K} x_i(t)$$

Intrinsic Mode Function, $x_i(t)$

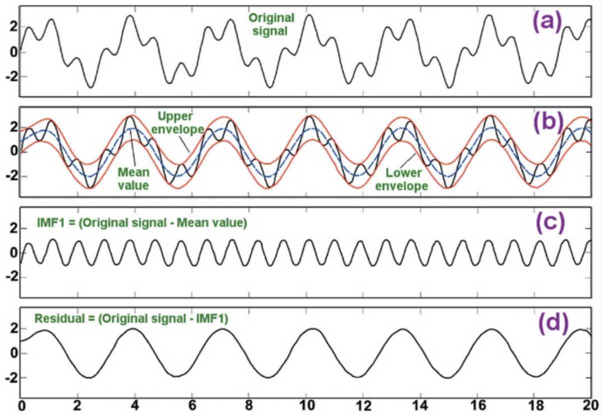


Multivariate EMD [2]

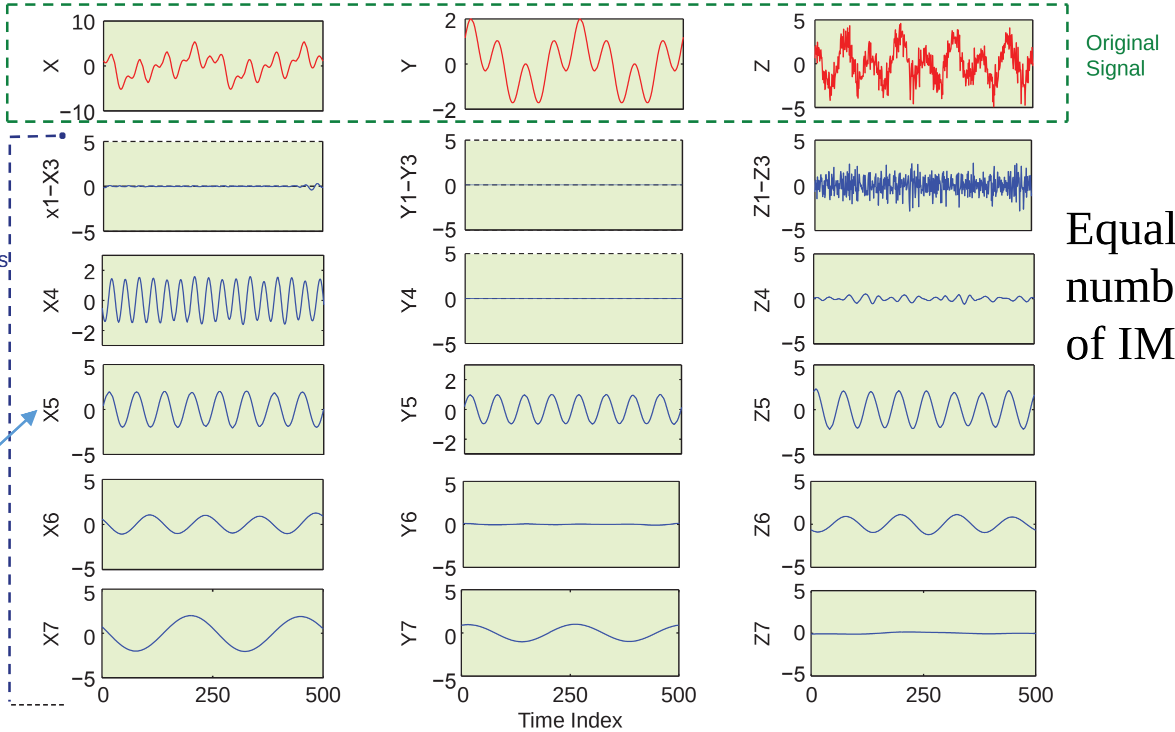
Trivariate signal s = [X,Y,Z]

X = [8 Hz + 16 Hz],
Y = [8 Hz + 4 Hz],
Z = [8 Hz + 2 Hz + noise].

Common Mode



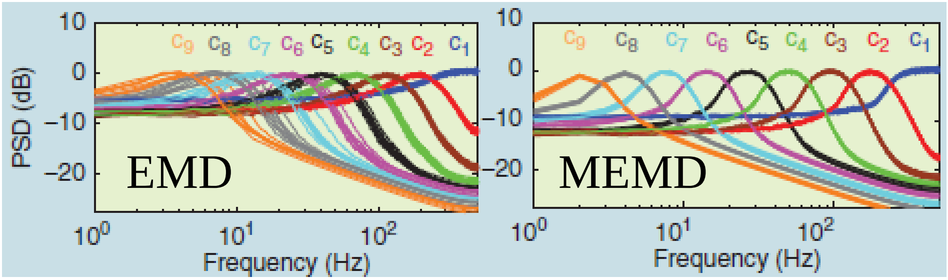
Hansan [3]



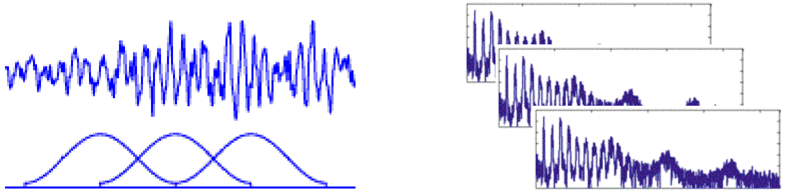
Equal number of IMFs

IMFs: Intrinsic Mode Functions

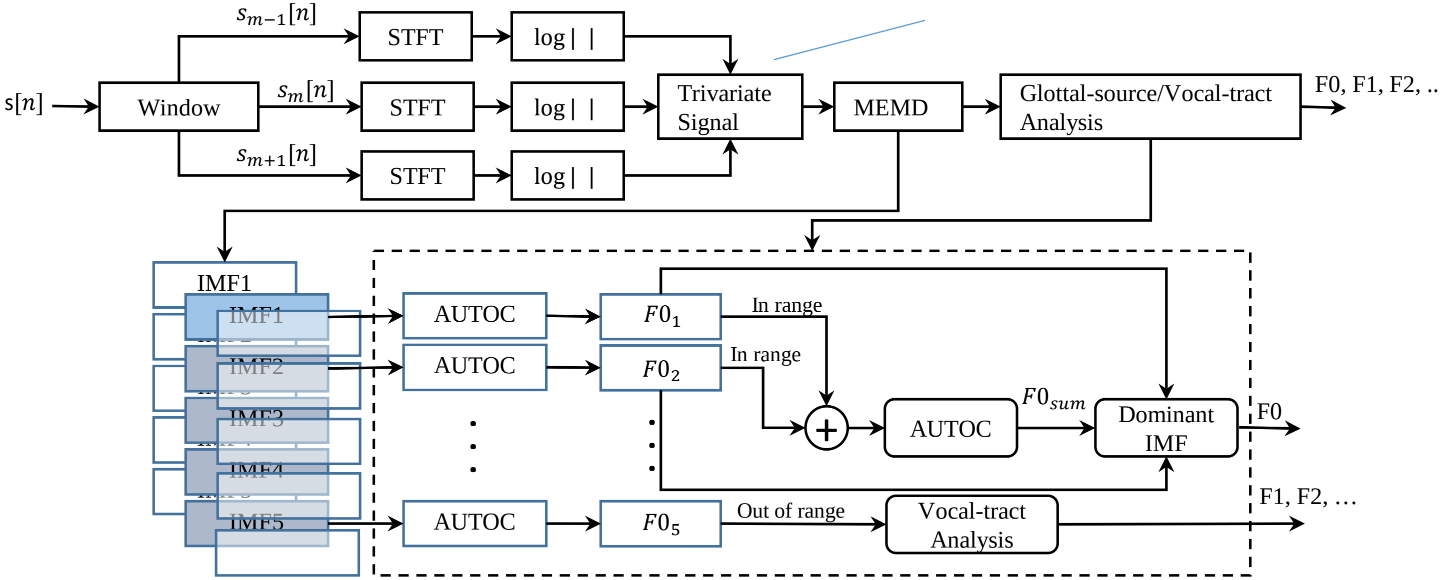
Band of IMF



Proposed Method

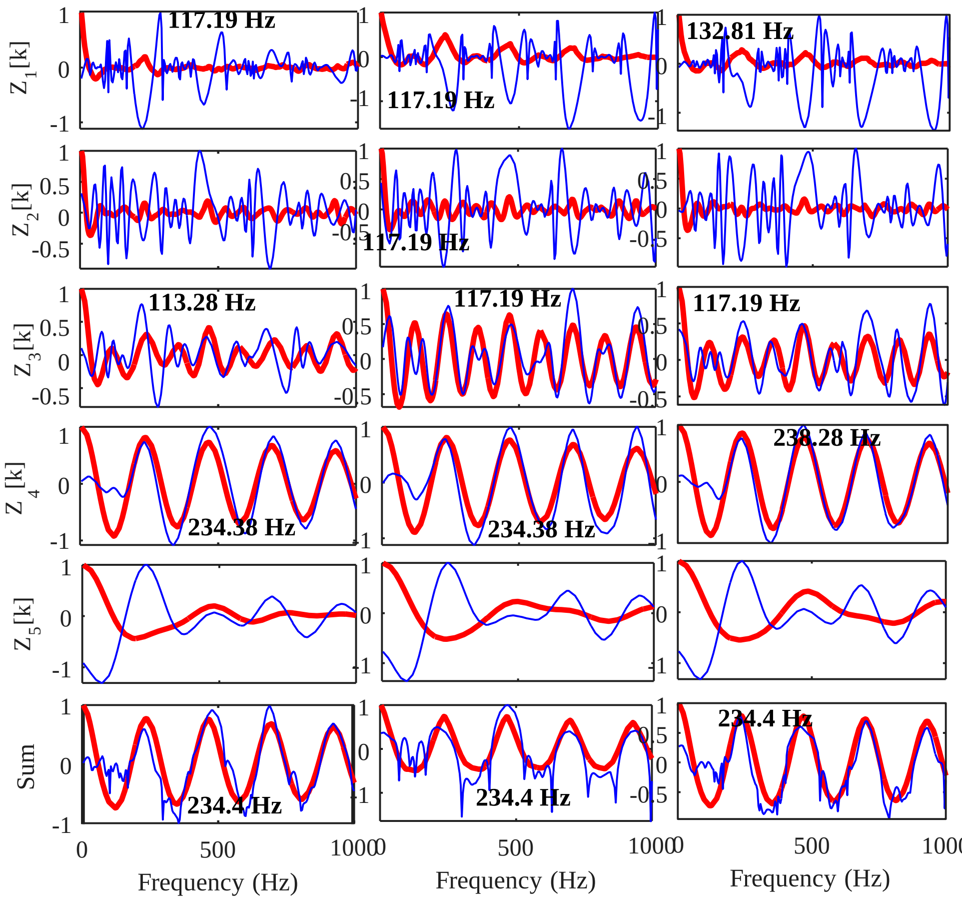


$\log|\mathbf{S}[k]| = [\log|S_1[k]|, \log|S_2[k]|, \log|S_3[k]|]$



Divide IMFs into two groups of glottal-source and vocal-tract using autocorrelation (AUTOC).

$$\log|S[k]| = \log|E[k]| + \log|H[k]| = \sum_{i=1}^{K} Z_i[k] = \underbrace{\sum_{i=1}^{M} Z_i[k]}_{glottal-source} + \underbrace{\sum_{M+1}^{K} Z_i[k]}_{vocal-tract}$$

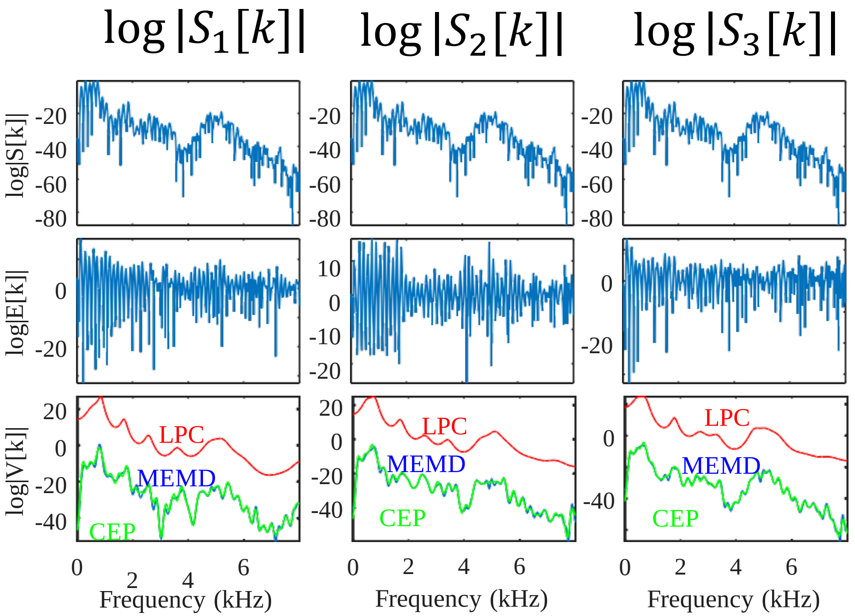


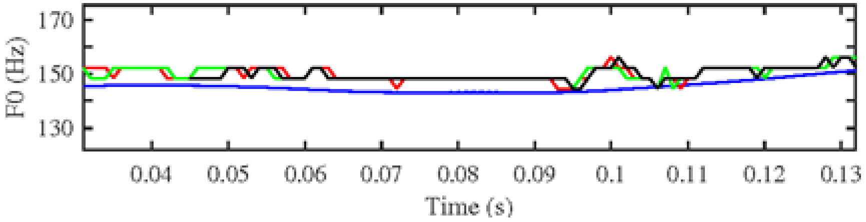
IMFs and their autocorrelation (AUTOC)

If the first peak of AUTOC is between 85 – 255 Hz (normal range of F0) then that IMF are considered as of glottal-source.

Common mode alignment in $Z_4[k]$

Evaluation & Results



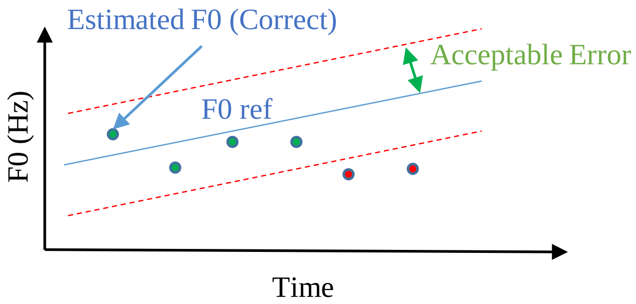


Itakura-Saito Distance

$$D_{IS}(P(\omega), \hat{P}(\omega)) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{P(\omega)}{\hat{P}(\omega)} - \log\frac{P(\omega)}{\hat{P}(\omega)} - 1\right] d\omega$$

- Glottal-Source
  - F0 estimation

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



- Vocal-tract
  - Formants (F1, F2)
  - Shape
- Peak detection
- Correlation
- Spectral Distance

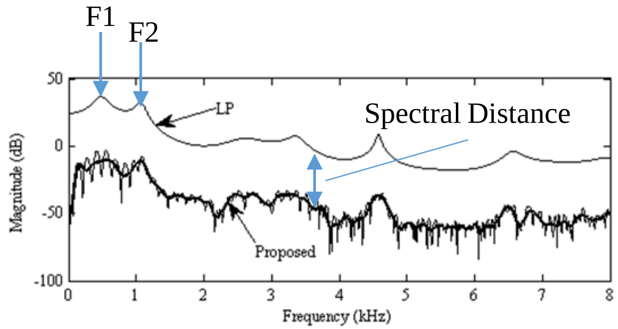


Table 1: Correct rate (%) of F0 estimation using proposed method compared with those obtained by linear prediction (LP) and cepstrum (CEP)

Vowel	LP	CEP	Proposed
/AA/	94.12	92.87	93.88
/IY/	87.24	89.64	90.13
/UW/	95.65	86.25	92.02
/EY/	92.08	90.39	92.61
/OW/	91.52	93.79	90.46

Table 2: Average formant frequencies (kHz) and spectral distance.

Vowel	Method	F1	F2	Correlation		D_IS	
				E-C	E-L	E-C	E-L
/AA/	LP	0.74	1.45	0.98	0.95	0.06	96.03
	CEP	0.76	1.43				
	Proposed	0.76	1.47				
/IY/	LP	0.37	2.20	0.99	0.94	0.03	124.76
	CEP	0.37	2.22				
	Proposed	0.36	2.22				
/UW/	LP	0.40	1.35	0.98	0.93	0.05	253.02
	CEP	0.39	1.34				
	Proposed	0.35	1.34				
/EY/	LP	0.47	2.05	0.99	0.95	0.03	83.73
	CEP	0.45	2.06				
	Proposed	0.46	2.06				
/OW/	LP	0.58	1.38	0.99	0.92	0.05	174.76
	CEP	0.58	1.31				
	Proposed	0.57	1.34				

Conclusion

- The proposed method can automatically separate glottal-source and vocal-tract filter.
- The correct rate of estimated F0 obtained by the proposed method is as good as those obtained by LP-based and cepstrum-based methods.
- The estimated formants (F1 and F2) using proposed method are equivalent to those obtained by LP-based and cepstrum-based methods.
- The shape of spectral envelop is most similar to that obtained by cepstrum-based methods.

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.